D3.7 DE4A Machine Learning Algorithms

Document Identification			
Status	Final	Due Date	30/09/2022
Version	1.0	Submission Date	03/10/2022

Related WP	WP3	Document Reference	D3.7
Related Deliverable(s)	D2.4, D3.2, D3.4, D3.6, D4.1, D4.5, D4.9, D5.3	Dissemination Level (*)	PU
Lead Participant	SU	Lead Author	T. Karunaratne (SU)
Contributors	A.-R. Guzman (MTFP-SGAD), Slavko Žitnik, (SI-MPA), E. Kapantai, I. Konstantinidis (IHU), C.Berberidis (IHU)	Reviewers	Damijan Novak (UM) Miguel Correia (INESC-ID)

Keywords:
Data, log data, Machine learning, Artificial Intelligence, visualisation, Semantic Assets, Core Vocabularies, Semantic interoperability, Canonical Evidences, Evidence Exchange, Semantic Model, Public Procedures, Information Exchange

Document Information

List of Contributors	
Name	Partner
Thashmee Karunaratne	SU
Christos Berberidis	IHU
Eleni Kapantai	
Ioannis Konstantinidis	
Ana-Rosa Guzman	MTPF-SGAD
Slavko Žitnik	SI-MPA

Document History			
Version	Date	Change editors	Changes
0.1	01/04/2022	T. Karunaratne (SU)	Initial version of document
0.2	19/09/2022	T. Karunaratne (SU), A. R. Guzman (SGAD), S. Žitnik (SI-MPA), E. Kapantai, I. Konstantinidis (IHU)	Content writing of all chapters - in an online document
0.3	21/09/2022	T. Karunaratne (SU)	Finalising for internal review
0.4	21/09/2022	E. Kapantai (IHU)	Minor comments
0.5	23/09/2022	D. Novak (UM) Miguel Correia (INESC-ID)	Internal review
0.6	28/09/2022	T. Karunaratne (SU)	Final version submitted to the coordinator
0.7	28/09/2022	Julia Wells (Atos)	Revision for submission
1.0	30/09/2022	Ana Piñuela (Atos)	Final version for submission

Quality Control		
Role	Who (Partner short name)	Approval Date
Deliverable leader	Thashmee Karunaratne (SU)	28/09/2022
Quality manager	Julia Wells (ATOS)	28/09/2022
Project Coordinator	Ana Piñuela Marcos (ATOS)	30/09/2022

Table of Contents

List of Tables

List of Figures

List of Acronyms

Abbreviation / acronym	Description
API	Application Programming Interface
AI	Artificial Intelligence
BB	Building Block
CEF	Connecting Europe Facility
CPOV	Core Public Organization Vocabulary
CPSV	Core Public Services Vocabulary
DBA	Doing Business Abroad – one of the three DE4A pilots
DC	Data Consumer
DE	Data Evaluator
DE4A	Digital Europe for All
DM	Data Mining
DO	Data Owner
DP	Data Provider
DR	Data Requestor
DSD	Data Services Directory
DT	Data Transferor
EC	European Commission
EDCI	Europass Digital Credentials Infrastructure
EDM	Exchange Data Model
EDU	Educational competent authority
eID	Electronic Identification
EIF	European Interoperability Framework
ESL	Evidence Service Locator
IAL	Issuing Authority locator
IDK	Information Desk
IEM	Information Exchange Model
IHU	International Hellenic University
IOP	Interoperability
ISA2	Interoperability solutions for public administrations, businesses and citizens
ISCED-F	International Standard Classification of Education-Fields
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
MA	Moving Abroad – one of the three DE4A pilots
MOR	Multilingual Ontology Repository
MVP	Minimum Viable Product
NLP	Natural Language Processing
NUTS	Nomenclature of Territorial Units for Statistics
OASIS	Organization for the Advancement of Structured Information Standards

Abbreviation / acronym	Description
OOP	Once-Only Principle
OOTS	Once-Only Technical system
OWL	Web Ontology Language
PKI	Public Key Infrastructure
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
REST	Representational state transfer
SDG	Single Digital Gateway
SLR	Systematic Literature Review [1]
SKOS	Simple Knowledge Organization System
SPARQL	SPARQL Protocol and RDF Query Language
TM	Text Mining
UC	Use Case
UML	Unified Modelling Language
UI	User Interface
W3C	World Wide Web Consortium
WP	Work Package
XML	Extensible Markup Language
XSD	XML Schema

Glossary

Abbreviation / acronym	Description
Application Profile	An application profile (AP), as yet another group of assets within the 'models' category, describes how a standard is to be applied in a particular domain or application. Standards typically do not contain constraints such as cardinality; these constraints are defined in the application profile. An application profile only applies to the specified domain [2].
Code lists	Predefined set of terms from which some statistical coded concepts take their values [3].
Canonical Evidence Type	Evidence type defined by an agreement on the fact proved and the information provided, along with a structured data model for the common set of attributes [4]. Used to provide a common classification of domestic evidence types and a semantic interoperability agreement for their contents to remove linguistic and semantic barriers.
Canonical Evidence	Piece of evidence issued according to a certain canonical evidence type by an issuing authority, who guarantees that its information is consistent with the information provided by the corresponding lawful domestic evidence.
Canonical Event Catalogue	Catalogue of events that change the contents of a base registry, according to a semantic agreement.
Controlled Vocabulary	A consistent way to describe data. They are standardized and organized arrangements of words and phrases presented as alphabetical lists of terms or as thesauri and taxonomies with a hierarchical structure of broader and narrower terms [4].
Criteria	Procedural requirements as conditions to be met and used as a basis for making judgements or decisions in the procedure.
Data Model	A data model is a collection of entities, their properties and the relationships among them, which aims at formally representing a domain, a concept or a real-world thing. It includes core vocabularies [6].
Ontology	An ontology is a formal, explicit specification of a shared conceptualisation. In computer science and information science, an ontology encompasses a representation, formal naming and definition of the categories, properties and relations between the concepts, data and entities that substantiate one, many or all domains of discourse [7].
Procedure	A sequence of actions that must be taken by users to satisfy the requirements, or to obtain from a competent authority a decision, in order to be able to exercise their rights as referred to in point (a) of Article 2(2) [8].
Provision	Availability of a service offered by a specific issuing authority to provide a specific canonical evidence type (evidence provision) or the subscription to a specific Canonical Event Catalogue (subscription provision).
Scenario	One typical way in which a system is used or in which a user carries out some activity.
Semantic Asset	A specific type of standard which involves highly reusable metadata (e.g., XML Schema, generic data models) and/or reference data (e.g. code lists, taxonomies, dictionaries, vocabularies).

Abbreviation / acronym	Description
Semantic Component	A component (e.g. Information Desk, Information Exchange Model) of the semantic interoperability framework that uses semantic assets to perform certain functionalities.
Semantic Interoperability Framework	A framework that consists of semantic components and the related semantic assets to facilitate cross-border exchange of evidences.
Taxonomy	A systematic arrangement in groups or categories of concepts according to established criteria [9].
Use case	A specification of one type of interaction with a system. One use case may involve several scenarios (usually a main success scenario and alternative scenarios).
User	Anyone who is a citizen of the EU, a natural person residing in a Member State or a legal person having their registered office in a Member State, and who accesses the information, the procedures, or the assistance or problem-solving services, referred to in Article 2(2), through the gateway [10].
Vocabulary	A collection of terms for a particular purpose. Vocabularies can range from simple ones, such as the widely used RDF schema, FOAF and DCMI element sets, to complex vocabularies with thousands of terms, such as those used in healthcare to describe symptoms, diseases and treatments. Vocabularies play a very important role in linked data, specifically to help with data integration. For example, metadata vocabulary. The use of this term overlaps with that of 'ontology' [11].
XML Schema	An XML schema is a description of a type of XML document, typically expressed in terms of constraints on the structure and content of documents of that type, above and beyond the basic syntactical constraints imposed by XML itself [12].

Executive Summary

The implementation of the Once-Only Principle (OOP) for public services at European level faces a great challenge: semantic interoperability. To a large extent, cross-border semantic interoperability has been addressed by developing common data models and formats for the information to exchange. This, however, is not a simple task and need to solve many other issues in order to achieve semantic interoperability in this context. Could Artificial Intelligence (AI) help in providing a coherent semantic interoperability for cross-border public services?

Many studies on how AI could help to develop e-government can be found in the literature, but none of these studies refers to cross-border information exchange. Application of Machine Learning (ML) regarding once-only information exchange is yet to be introduced into academic research in the field. Cross-border information exchange through the OOP for public services is mainly concerned with personal data, which lawfully require strict data protection measures, which limit their analysis using any big data technology. Only metadata of the exchange regarding competent authorities, evidence types and data services data could possibly be openly analysed. But these metadata are neither large enough nor are suitable to successfully apply ML algorithms, even with the techniques tailored for small data sets, due to the meta data structure, incompleteness (due to practical reasons for e.g., some data in some registries may not be available before a particular year or so), and the nature of the data exchanged.

To understand how useful and reliable the technical system is, how to strengthen its resources, how to prevent errors, etc., the performance of a system should be monitored and analysed on the basis of a sound logging mechanism. The DE4A Once-Only technical system provides a logging mechanism embedded in the distributed common components that implement the exchanges, which are operated by several stakeholders at different Member States: Authority Agents and Connectors. The data logged by these common components along with the metadata of objects and participants of the exchange, which are provided by the Information Desktop, are data sources for metrics proposed to measure the performance of our system. Semantic agreements have been put in place for both description of metrics and data sources to allow their automatic processing. In this way, besides traditional data analysis, ML can be used to identify clusters of anomalies in the system to detect and alert on significant incidents and to help on failure prevention and system performance.

ML can also be applied to learn from existing semantic assets, usually described by text documents. As a result recommendations can be made for the semantic experts that are designing new ones, such as common data models for information to exchange. Conversely, semantic experts can help ML to both understand the precise meaning of concepts within public administration context at each Member State and identify/discard equivalences or relationships. In the context of public services, AI, specifically text mining, can be a major contribution to process the huge number of usually changing legal texts that provides governments' rules for processing administrative procedures, issuing evidence, and distributing responsibilities among public entities. However, eGovernment and interoperability require organisational catalogues where such information is represented as known structured data. AI could extract knowledge from legal texts and semantic asset textual descriptions to automatically feed and maintain organisational catalogues and reusable semantic asset catalogues to be automatically processed.

In this deliverable we propose AI applications to process legal text and documentation of reusable semantic assets, combining multilingualism, low-resource languages and different domains of documents. Finally, we propose the use of Chat bots to help European citizens to locate public services and proper evidence, thus helping them in using the system. The scope of the DE4A project limits the grounds to develop AI tools for the aforementioned context. Furthermore, the scarcity of academic evidences for such attempts limits us only to propose possible avenues for application of ML

techniques, without any proof of the concepts. However, when and wherever the data is available, the suggestions provided by this deliverable sets up the basic grounds for future application and the practical use of AI in the implementation of the OOP for public services within and across borders.

1 Introduction

Artificial Intelligence (AI), specifically Machine Learning (ML) consists of advanced technologies that utilizes data efficiently for drawing conclusions on the processes the data are subjected to. These advanced technologies are gradually becoming popular in the organisations of the public sector, since these technologies enable enhancing the performance of the systems and the services provided by the organisations. Governments in different parts of the world have already attempted implementation of technologies based on AI techniques to digitalize and improve internal E-Government services. Examples for such services include evidence based public policy making, enhanced delivery mechanisms, promoting information provision to citizens and businesses and so forth. New advanced applications powered by AI will be prevalent in organizations to automatize processes and manage information and knowledge according to contemporary literature. Evidence also show that AI can make a considerable impact on decision-making process as well as detection of the anomalies in the service provision process and reaction to changes in the environment [13]. Personalisation of e-government services let government-citizen interaction to be efficient and real time. It allows increased interoperability by data driven information retrieval and processing of sheer amount of data for detecting patterns, discovering new solutions through dynamic models and simulation in real time.

The cross-border and cross-sector public services provisions enabled by the Once Only Principle ideally extends the technical complexity of the e-government services. However, AI (and its variants such as ML, text mining, data mining, visualisations etc.), opens up a strong potential in reducing the complexity and legal, business and technical challenges in automatic information exchange as the current research shows. It is also important to note that, although our interest is in the semantic layer of the cross-border public service provision, our preliminary attempts of exploring the AI in e-Government landscape failed to spot in any significant resource directed towards addressing semantic interoperability. Such a gap leaves extensive problems, since semantics sets up the structure enabling communication between systems.

On the other hand, even though the cross-border automated public service provisioning under OOP is the main focus of DE4A, rather than the internal automated e-government and its approaches, the current applications and practices of automated e-government solutions powered by advanced technologies synergise the cross-border counterpart up to a significant extent. Thus, understanding the state-of-the-art e-Government powered by AI, may provide the ground resources and basic understanding for public service extension across-borders. However, the limitation of the previous examples of semantic interoperability related attempts and the scarcity of data from the semantic components developed in DE4A and elsewhere limit us to only to provide speculations and suggestions for potential AI possibilities applied on synthetic data.

1.1 Purpose of the document

This document presents the outcomes of the activities WP3 carried out in conjunction with application of ML techniques to enhance the performance of the semantic interoperability layer of the Once Only Technical System (OOTS) the DE4A project implemented. The task T3.4 Machine learning applications presumably looked upon achieving two goals 1) investigation of potential machine learning solutions that could efficiently enhance semantic interoperability between the integrated e-services, and 2) methodologies for automated analysis of usage data derived from piloting of the use cases, to establish semantics upon the services utilized. Limited by the unavailability of data from the piloting use cases, this task and hence the deliverable adjusted the focus towards investigating the state-of-the -art of ML and AI applications for semantic interoperability, and conceptualising potential machine applications.

In this document AI, ML, and other text, big data and analytical methods are collectively referred as AI, or vice versa.

1.2 Structure of the document

This deliverable is structured to provide firstly, an overall understanding of the state-of-the-art methods and practices of advanced technologies using a Systematic Literature Review (SLR), and secondly, an overall understanding of the data available in static and dynamic forms. Based on the contemporary knowledge, it also illustrates some proofs of concepts of using ML in the domain, and also some other conceptual proposals for prospective application of advanced technology for increased efficiency of the service provisions.

- Chapter 2 consists of the scientific literature study SLR that provides the state-of-the-art of the digital government.
- Chapter 3 contains a complete overview of data that is and can be available for and as a result of using OOTS. This chapter also gives an overview of how visualisation dashboards can be created from these data.
- Chapter 4 conceptualise the ML applications that support enhancing some of the services and smoothening processes. Some of the solutions precented in this chapter are proof-of-concepts that demonstrates the advancements, but some other applications are at the concept level with argumentations of how these concepts are promising.
- Chapter 5 finally summarises the outcome of the whole document.

2 Machine learning applications for cross-border public service

In this chapter the existing efforts for application of ML for the public services are systematically analysed. Hence a systematic literature study is conducted to investigate systematically the landscape of advanced technologies in cross-border public service provision.

As mentioned in the previous section, the intension of the Systematic Literature Review (SLR) is to find the AI landscape of e-Government irrespective of internal or cross-border. We believe, such knowledge will allow understanding the semantic scope of the e-Government, and the level of application of advanced technologies for enhancing semantic interoperability.

2.1 Systematic literature study for the state-of-the-art

2.1.1 SLR methodology

The systematic review is "*a review of the research literature whose aim is to arrive at a conclusion about the state of knowledge on a topic based on a rigorous and unbiased overview of all the research that has been undertaken on that topic*" [14].

In this systematic literature review, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework [1] is applied. PRISMA outlines four steps with which to conduct the systematic review (c.f. Figure 1):

1. Identification of records through peer-reviewed publications database searching and other relevant sources.
2. Screening of the identified records including removal of duplicates.
3. Eligibility assessment of full-text articles.
4. Included studies in qualitative synthesis and quantitative synthesis (meta-analysis).

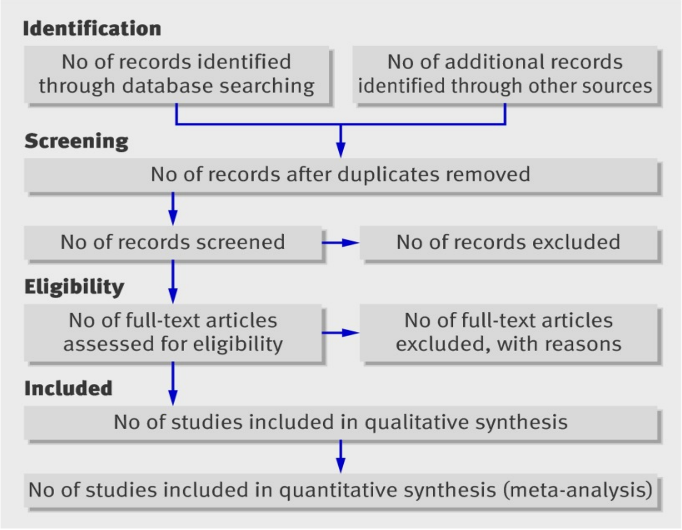

Identification
No of records identified through database searching
No of additional records identified through other sources
Screening
No of records after duplicates removed
No of records screened
No of records excluded
Eligibility
No of full-text articles assessed for eligibility
No of full-text articles excluded, with reasons
Included
No of studies included in qualitative synthesis
No of studies included in quantitative synthesis (meta-analysis)

Figure 1: Process of creating the literature corpus for the SLR [1]Article databases

For this SLR we consider the databases in Table 1, due to their multi-disciplinarity as they cover journal articles from all subject areas and includes results from a large selection of databases.

Table 1: List of Databases

Database	Description	Links if any
EBSCO Discovery Services	"EBSCO Discovery Service (EDS) provides journal articles from all subject areas" (Stockholm University)	https://www.ebsco.com/products/ebsco-discovery-service
Scopus	"Scopus is the largest abstract and citation database of peer-reviewed literature – scientific journals, books and conference proceedings" (blog.scopus.com)	https://www.scopus.com/
Web of Science Core Collection	"The Web of Science is a bibliographical database of scholarly articles from 22,000 peer-reviewed journals worldwide." (Eui.eu)	https://www.webofscience.com/
Google Scholar	"Google Scholar is a Web search engine that specifically indexes scholarly literature and academic resources."	https://scholar.google.com/
PubMed	"PubMed is the number one resource for anyone looking for literature in medicine or biological sciences. PubMed stores abstracts and bibliographic details of more than 30 million papers and provides full text links to the publisher sites or links to the free PDF on PubMed Central (PMC)."	https://pubmed.ncbi.nlm.nih.gov/
IEEE Xplore	"IEEE Xplore is the leading academic database in the field of engineering and computer science. It's not only journal articles, but also conference papers, standards and books that can be search for."	https://ieeexplore.ieee.org/Xplore/home.jsp
ScienceDirect	"ScienceDirect is the gateway to the millions of academic articles published by Elsevier. 2,500 journals and more than 40,000 e-books can be searched via a single interface."	https://www.sciencedirect.com/
Directory of Open Access Journals	"The DOAJ is very special academic database since all the articles indexed are open access and can be accessed freely of charge."	https://doaj.org/
JSTOR	"JSTOR is another great resource to find research papers. Any article published before 1924 in the United States is available for free and JSTOR also offers scholarships for independent researchers."	https://www.jstor.org/

2.1.2 Concepts for searching

To capture the relevant articles systematically and also to facilitate reproducibility, search protocols are required. These protocols typically consist of keywords connected by Boolean operators. To identify correct combinations of keywords, the key concepts are to be identified as the first step. Table 2 below gives concept descriptions leading the search protocol creation. We also have tried OOTS specific keywords with ML and AI keywords, with an ambition to see the DE4A specific interest, without any hits. While pointing out how ML in OOTS domain is underrepresented in contemporary research, we focussed on a wider scope of e-government, as shown below.

Table 2: Key concepts

Concept 1	Concept 2	Concept 3
Public Administration	Digital Governance	Machine Learning
Cross-domain	e- Government	Artificial Intelligence
Cross-border	Digital Government	Semantic web

2.1.3 SLR Protocols and inclusion and exclusion criteria

Based on the concepts identified in Table 2, the keywords are created. These keywords are combined to create the search protocols. These protocols also include the criteria, for, e.g., we look for only peer-reviewed publications in English, and the search terms present in the titles and abstracts. The complete structure of the protocols are as follows:

- **Protocol 1**: “Public Administration” AND “Digital Governance” AND (“Machine Learning” OR “Artificial Intelligence”)
  - Criteria 1 inclusion: 1, 2, 3
  - Criteria 2 exclusion: 4, 5, 6
- **Protocol 2**: (“Cross-domain” OR “Cross-border”) AND (“e- Government” OR “Digital Government”) AND “Semantic web”
  - Criteria 1 inclusion: 1, 2, 3
  - Criteria 2 exclusion: 4, 5, 6

The Inclusion criteria is as follows:

Table 3: Criteria of inclusion of searched papers

	Criteria	Motivation
1	Peer-reviewed Academic publications (including journal articles, conference papers and books as well as grey literature such as theses and doctoral dissertations)	This SLR is limited to academic publications in order to set and maintain a consistently high quality globally accepted standard from which the results are systematically proven.
2	Written in English	Only focuses on academic publications written in English or officially translated to the English language.
3	Published from 2012 to 2022	This research will draw from the last decade of publications in order to draw the latest best practices

We followed the below mentioned exclusion criteria as well:

Table 4: Criteria of exclusion of searched papers

	Criteria	Motivation
4	Published before 2012	Many advanced technologies are adapted in public service provisions during the past decade.
5	Published in other languages	We do not examine articles written in other languages than English.
6	Does not include the keywords in the title or abstract	

Table 5 shows the total hits from different publication databases based on the abovementioned selection queries.

Table 5: Results of systematic search of databases

	Protocol 1	Protocol 1 + Filtering Criteria 1 and 2	Protocol 2	Protocol 2 + Filtering Criteria 1 and 2
EBSCO Discovery Service (EDS)	108	46	8	3
Scopus	238	207	66	46
Web of Science	44	40	5	4
Google Scholar*	135.000 + 150.000	60	1.200 + 13.500 + 2.800 + 13.600	40
PubMed*	12+4	16	0 + 0 + 0 + 0	0
IEEE Xplore	1	1	0	0
ScienceDirect	34	33	19	16
Directory of Open Access Journals **	/	/	/	/
JSTOR	9	8	3	1
Total	*****450**	**411**	*****101**	**110**

* Search engine does not allow precise filtering and many results are returned. After the protocol search combinations (two for Protocol 1 and four for Protocol 2) we check first 30 results and select only those papers whose title seemed relevant.

** Papers already included in EBSCO database.

*** Google scholar not taken into account.

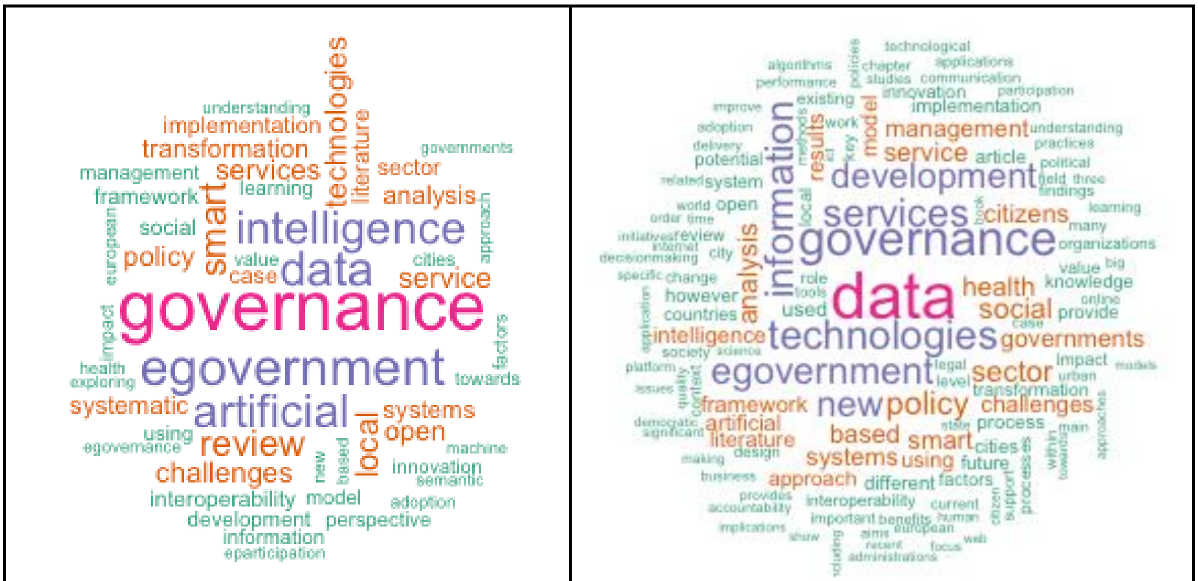
Removing duplicates across databases and fine screening resulted in 399 prospective articles.

2.2 Automatic analysis of screened publications

The set of papers qualified through the step of Inclusion and exclusion (i.e., the step of screening in PRISMA) is, in principle, in the broader subject area, although it was subjected to further trimming due to the size of the corpus (the step of Eligibility in PRISMA framework). 399 papers were kept after the Screening step. To understand the scope of the applicability (or discussion) about advanced technologies in the contemporary research, an overall visualisation and analysis was conducted using automated literature analysis by text mining methods.

2.2.1 Word frequency analysis using word clouds on topics and abstracts

Automated literature analysis allows getting a wider aspect of the trends in the field of concern with less effort compared to its manual counterpart of traditional SLR. Hence, an automatic analysis was conducted across the article corpus of 399 articles. Firstly, a visualisation of words in titles and abstract was executed to get an overview of the focus. Figure 2 represent the word cloud of the titles (left), and abstracts (right). The summary of words in the word cloud is obtained based on standard text mining methods such as, words extraction, stop word removal, and by term document metrics. In creating the word cloud the most common words for the domain such as “research", "study", "government", "public", "administration", "digital", "e", "technology", "paper" is also removed in addition to the standard English stop words and non-English words.



(wordcloud(min.freq = 10, max.words=200)

wordcloud(min.freq = 50, max.words=200)

Figure 2: Word cloud of the titles (left), and abstracts (right) of 399 articles

The word clouds of titles and abstracts do not significantly deviate from each other, and shows homogeneity in the top key words in them. Naturally, data and governance are dominant terms. Titles have “artificial Intelligence” and “eGovernment” also as prominent words, but surprisingly terms related to cross-border transactions are neither appearing in titles nor in abstracts. Therefore, as our pre-assumption, attempts to apply AI and Machine learning (including analytics and reasoning) on data from harmonised public service are yet to appear in the academic community of the subject.

2.2.2 Mining the document corpus: Cluster analysis for topic filtering

In a typical SLR, the first round of document selection is based on the content in the abstracts. The automatic analysis of the document corpus hence includes the abstracts as instances (individual documents). Text mining of documents allow classifying the documents based on the texts in them. When the documents are unstructured, for example, documents are unformatted and in a natural language (English), clustering approaches are used to classify them. Cluster analysis is useful for identifying different research lines (trends) in the focussed field based on the contents of the documents (abstracts of selected papers). Typical cluster analysis algorithms identify the clusters using standard methods such as hierarchical or k-means clustering. Hierarchical clustering provides a grouping of documents based on the word frequencies in each group and the distance between the most frequent words in these groups. In this analysis we have used agglomerative (bottom-up) clustering method to build the hierarchical topic tree with 4 clusters in Figure 3(left). In contrast to hierarchical, partitional clustering methods such as k-means algorithm, result in groups of documents that are best disjoint from each other. The advantage of applying such an algorithm on a document corpus is that it allows identifying how different the themes the documents represent. In the context of the SLR k-means clustering visualizes the dispersion of words from the canter of the cluster, and hence showing how close the documents in each cluster to each other. In this study we applied k-means algorithm as implemented in R statistical software package; version 2022.07.1 Build 554 of RStudio. We have been using the term document frequency of the abstracts, with parameters of k (number of clusters) = 4 and nstart (starting number of cluster configurations) = 100. Figure 3 (right) shows the visual outcome of k-mean clustering. The number of clusters (4) was a parameter estimated in trial-and error fashion.

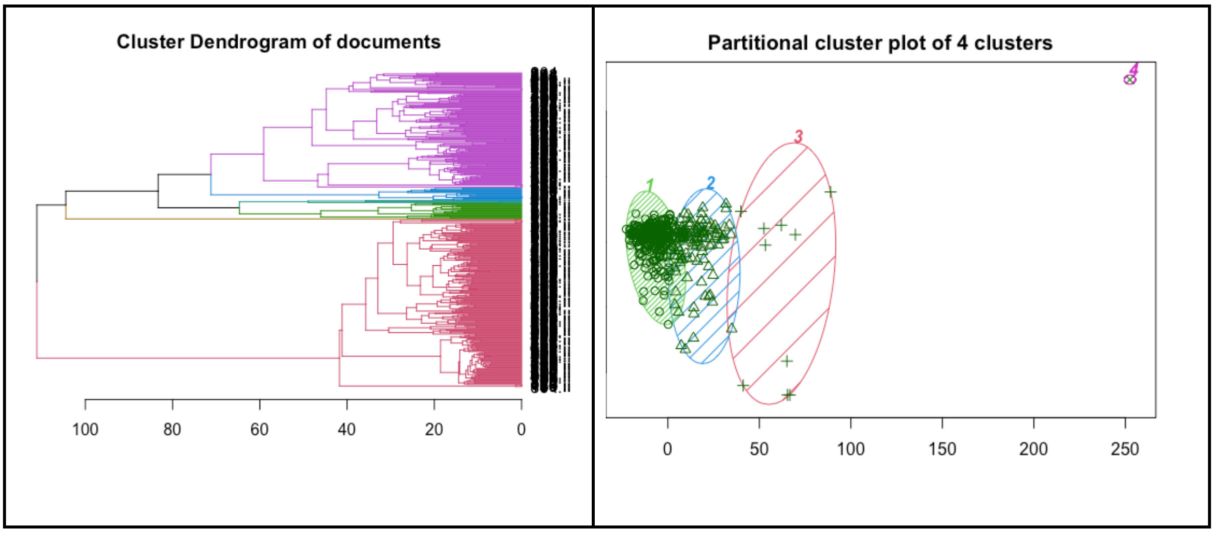


Figure 3: Cluster dendrogram (left) and K-means clustering (right)

The most frequent 10 words of each cluster is as follows:

Table 6: 10 most frequent words in each cluster

Clusters and freq.	1	2	3	4	5	6	7	8	9	10
Cluster_1										
Term	govern	data	develop	servic	inform	technolog	system	process	sector	manag
Frequency	116	86	84	78	62	59	51	45	44	42
Cluster_2										
Term	govern	servic	data	technolog	develop	system	model	inform	manag	provid
Frequency	197	167	150	127	120	81	75	71	71	69
Cluster_3										
Term	govern	data	servic	develop	polici	inform	technolog	provid	system	citizen
Frequency	191	143	143	121	114	97	88	84	80	75
Cluster_4										
Term	govern	technolog	data	develop	sector	servic	inform	health	value	polici
Frequency	122	79	75	73	53	52	51	50	47	46

The word distributions of the 4 clusters in k-means are 1) 308, 2) 112, 3) 10 and 4) 1, with the group dispersion:

Within cluster sum of squares by cluster:

41526.9, 6269576.6, 4192672.91   0.00

between_SS / total_SS = 80.6 %

This means the variations of the first 3 clusters are high and inter-group variations less significant. According to the visualisations of both clustering approaches, a heavy imbalance of the word distribution is observed, which is obvious from the table of cluster frequencies also. The words in (abstracts) are overlapping in a significantly high degree. One document however stands out from others, which is apparently an outlier.

Therefore ultimately it can be concluded the abstracts in the document corpus use sets of words that are not distant from each other; in other words, the focus or the field of research is concentrated around similar contexts.

2.2.3 Topic Modelling

It is worth seeing if significantly different topics could be identified even with a similar set of words, since it is of SLR's interest to find the different focusses of the document corpus. A topic modelling approach is followed in this regard. Topic mining algorithms mine hidden semantic structures from unstructured text data. The main purpose of topic models is to discover the topics the bundle of documents represents, allowing to understand the overarching concepts the texts present. In this study we have applied standard methods for probabilistic topic modelling, such as Latent Dirichlet Allocation (LDA). LDA is a parameterised method of finding themes different documents can be grouped into, based on the words they contain and their semantic relation.

All in all the following terms are the most frequent words appearing across most of the documents: "govern" "servic" "develop" "citizen" "model" "digit" "inform" "particip" "factor" "valu" . Note that the terms are stemmed. In preparation of the documents for LDA, in addition to the pre-processing of documents for clustering as described in the previous section, the terms are stemmed using standard stemming function as implemented in R statistical software. LDA require parameter optimisation for increased accuracy of topics, but we remained within the standard default parameter set. Two variations tested to monitor the difference between the suggested topics 1) standard LDA, and 2) LDA with Gibbs sampling. Gibbs sampling assumes the underlying complex distribution in high dimension, based on the local dispersion of variables (and instances). In a complex field of 1000+ variables (words) and 399 instances (documents), Gibbs sampling could result in significantly accurate topics estimates.

We have used the following parameters in application of LDA algorithm:

```
#Set parameters for Gibbs sampling
burnin <- 4000
iter <- 2000
thin <- 500
seed <-list (2003,5,63,100001,765)
nstart <- 5
best <- TRUE
```

The standard LDA (Table 7 (left) and the LDA with Gibbs (Table 7(right)) returned topics with slightly different terms in it.

Table 7: Topic models with standard LDA (left) and LDA with Gibbs sampling (right)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	govern	servic	servic	govern	health
2	develop	technolog	govern	data	data
3	provid	model	smart	technolog	technolog
4	system	develop	digit	process	polici
5	polici	manag	citi	polici	develop
6	support	understand	system	develop	social

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	inform	servic	data	govern	polici
2	system	citizen	process	technolog	sector
3	implement	develop	manag	social	health
4	applic	model	approach	citi	intellig
5	present	framework	analysi	digit	artifici
6	challeng	valu	open	smart	impact

The topics derived from the two topic mining methods do not deviate significantly with each other as can be seen from the table above. Hence, based on the above outcomes, it can be concluded that the field of focus has trends around keywords 1) policy develop support systems 2) understand model services systems, 3) smart governance systems citizens 4) technology policy process develop, and 5) health data social policy technology. However, the automated analysis does not provide in depth understanding of the actual landscape of the trends, which paves the way for a manual analysis of a manually picked documents as described in the subsequent sections.

2.3 Selected articles for SLR

The 399 articles were further analysed to identify potential articles to include in the deeper analysis. The first round of abstract read resulted in 66 papers through. Further reading of the abstracts and skimming the whole articles resulted in a final corpus of 20 papers shown in Table 8, which we used for content analysis.

Table 8: Descriptions of selected papers

Paper ID	Description
#2	ALVAREZ, J. M., LABRA, J. E., CIFUENTES, F., ALOR-HÉRNANDEZ, G., SÁNCHEZ, C., & LUNA, J. A. G. (2012). TOWARDS A PAN-EUROPEAN E-PROCUREMENT PLATFORM TO AGGREGATE, PUBLISH AND SEARCH PUBLIC PROCUREMENT NOTICES POWERED BY LINKED OPEN DATA: THE MOLDEAS APPROACH. International Journal of Software Engineering & Knowledge Engineering, 22(3), 365–383. Business Source Premier.
#5	Avgerinos Loutsaris, M., Lachana, Z., Alexopoulos, C., & Charalabidis, Y. (2021). Legal Text Processing: Combing two legal ontological approaches through text mining. ACM International Conference Proceeding Series, 522–532. Scopus. https://doi.org/10.1145/3463677.3463730
#7	Blinova, N. V., Kirka, A. V., & Filimonov, D. A. (2021). Rational Bureaucracy 2.0: Public Administration in the Era of Artificial Intelligence. In E. G. Popkova, V. N. Ostrovskaya, & A. V. Bogoviz (Eds.), Socio-economic Systems: Paradigms for the Future (pp. 1679–1689). Springer International Publishing. https://doi.org/10.1007/978-3-030-56433-9_174
#9	Chen, Y.-C., Hu, L.-T., Tseng, K.-C., Juang, W.-J., & Chang, C.-K. (2019). Cross-boundary e-government systems: Determinants of performance. Government Information Quarterly, 36(3), 449–459. https://doi.org/10.1016/j.giq.2019.02.001
#15	Delgado, F., Hilera, J. R., Ruggia, R., Otón, S., & Amado-Salvatierra, H. R. (2021). Using microdata for international e-Government data exchange: The case of social security domain. Journal of Information Science, 47(3), 306–322. Library, Information Science & Technology Abstracts.
#18	Fragkou, P., Galiotou, E., & Matsakas, M. (2014). Enriching the e-GIF Ontology for an Improved Application of Linking Data Technologies to Greek Open Government Data. Procedia - Social and Behavioral Sciences, 147, 167–174. https://doi.org/10.1016/j.sbspro.2014.07.141
#19	Fullin Saldanha, D. M., Dias, C. N., & Guillaumon, S. (2022). Transparency and accountability in digital public services: Learning from the Brazilian cases. In GOVERNMENT INFORMATION QUARTERLY (Vol. 39, Issue 2). ELSEVIER INC. https://doi.org/10.1016/j.giq.2022.101680

Paper ID	Description
#20	Henman, P. (2020). Improving public services using artificial intelligence: Possibilities, pitfalls, governance. In ASIA PACIFIC JOURNAL OF PUBLIC ADMINISTRATION (Vol. 42, Issue 4, pp. 209–221). ROUTLEDGE JOURNALS, TAYLOR & FRANCIS LTD. https://doi.org/10.1080/23276665.2020.1816188
#21	HUIBREGTSE, S., & VERKAMMAN, J. (2021). The AI landscape for tax in Europe today. How Tax Authorities use AI and machine learning to facilitate their tax process. El Panorama Actual Del Uso de La Inteligencia Artificial Para La Fiscalización En Europa. Cómo Las Autoridades Fiscales Utilizan La Inteligencia Artificial y El Aprendizaje Automático Para Facilitar Sus Procedimientos Fiscales., 57(84), 191–232. edb.
#23	Kalogirou, V., Stasis, A., & Charalabidis, Y. (2022). Assessing and improving the National Interoperability Frameworks of European Union Member States: The case of Greece. Government Information Quarterly, 39(3), 101716. https://doi.org/10.1016/j.giq.2022.101716
#26	Leão, H. A. T., & Canedo, E. D. (2018). Best practices and methodologies to promote the digitization of public services citizen-driven: A systematic literature review. Information (Switzerland), 9(8). Scopus. https://doi.org/10.3390/info9080197
#28	Liva, G., Codagnone, C., Misuraca, G., Gineikyte, V., & Barcevicius, E. (2020). Exploring digital government transformation: A literature review. ACM International Conference Proceeding Series, 502–509. Scopus. https://doi.org/10.1145/3428502.3428578
#30	Margariti, V., Stamati, T., Anagnostopoulos, D., Nikolaidou, M., & Papastilianou, A. (2022). A holistic model for assessing organizational interoperability in public administration. Government Information Quarterly, 39(3), 101712. https://doi.org/10.1016/j.giq.2022.101712
#38	Patroumpas, K., Georgomanolis, N., Stratiotis, T., Alexakis, M., & Athanasiou, S. (2015). Exposing INSPIRE on the Semantic Web. Journal of Web Semantics, 35, 53–62. https://doi.org/10.1016/j.websem.2015.09.003
#46	Schmitz, P., Francesconi, E., Hajlaoui, N., & Batouche, B. (2018). PMKI: an European Commission action for the interoperability, maintainability and sustainability of Language Resources. In N. Calzolari, K. Choukri, C. Cieri, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga, S. Goggi, & H. Mazo (Eds.), PROCEEDINGS OF THE ELEVENTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2018) (pp. 2452–2455). EUROPEAN LANGUAGE RESOURCES ASSOC-ELRA.
#48	Shukair, G., Loutas, N., Peristeras, V., & Sklarß, S. (2013). Towards semantically interoperable metadata repositories: The Asset Description Metadata Schema. Computers in Industry, 64(1), 10–18. Scopus. https://doi.org/10.1016/j.compind.2012.09.003
#49	Sobrino-García, I. (2021). Artificial Intelligence Risks and Challenges in the Spanish Public Administration: An Exploratory Analysis through Expert Judgements. Administrative Sciences, 11(3), 102. https://doi.org/10.3390/admsci11030102
#51	Valle-Cruz, D., Alejandro Ruvalcaba-Gomez, E., Sandoval-Almazan, R., & Ignacio Criado, J. (2019). A Review of Artificial Intelligence in Government and its Potential from a Public Policy Perspective. Proceedings of the 20th Annual International Conference on Digital Government Research, 91–99. https://doi.org/10.1145/3325112.3325242

Paper ID	Description
#53	van Noordt, C., & Misuraca, G. (2022). Artificial intelligence for the public sector: Results of landscaping the use of AI in government across the European Union. Government Information Quarterly, 39(3), 101714. https://doi.org/10.1016/j.giq.2022.101714
#56	Waseem, A. A., Ahmed Shaikh, Z., & ur Rehman, A. (2016). A toolkit for prototype implementation of e-governance service system readiness assessment framework (Vol. 9752, p. 270). Scopus. https://doi.org/10.1007/978-3-319-39399-5_25

2.4 Summary of SLR outcomes

From the corpus of the documents a smaller set of documents were selected to manually investigate. The decision of which articles to be further investigate by reading the full texts are decided after reading the abstracts and skimming the full articles. The final selection of 20 papers were used for this deeper analysis. They are listed in the Table 8.

2.4.1 Attributes

We defined eight attributes based on which we reviewed the literature. The attributes are as follows:

- ***Year***: Publication year of the paper.
- ***Focus region***: Country or geographical area that was the focus of research.
- ***Public service***: Type of addressed public services (e.g., tax administration).
- ***Scope (size/type of study/use case or general)***: The size of the research that was performed. What amount of available data was used or what was the focus to produce the final outcome of the study.
- ***Methodology (list of methods used)***: The research methods used for conducting the study.
- ***Data (sources)***: Input data sources (e.g., existing literature, surveys, interviews) that were used for the study.
- ***Outcomes***: The main contributions or results of the study.
- ***Relevance to DE4A Semantics***: The main similarities to the work done within the DE4A project. These include similar approaches, methodologies or findings that might be useful in further implementation of the DE4A pilots.

In the following sections we separately review the main findings based on the results presented in Table 9 (note: articles are sorted ascending by the year of publication) below.

Table 9: Structured data extracted from the relevant papers

#	Year	Focus region	Public service	Scope (size/type of study/use case or general)	Methodology (list of methods used)	Data (sources)	Outcomes	Relevance to De4a Semantics
#2	2012	Whole EU	eProcurement	Processing and building semantic IS pipeline for 1,500 eProcurement notices per day.	1. Existing semantic vocabularies to RDF transformation 2. ETL of eProcurement notices into a modelled schema 3. Provisioning SPARQL-search endpoints	EU eProcurement notices	A new system/platform for data extraction from text and semantic search over the notices.	- Semantic data modelling - Usage of existing schemas and unification - Semantic search
#48	2013	General/EU	General	Use case prototype implementation for the selected data sources and feedback collection by the users.	1. Identify data suitable for repositories and objectives 2. Development of ADMS and mapping of metadata 3. Implementation of a prototype architecture 4. Evaluation feedback by 20 developers using the metadata	SEMIC.EU, Digitaliser, ESD toolkit service lists, XRepository	- Asset Description Metadata Schema (ADMS), a common metamodel for semantic interoperability assets	- Reusability of concepts - Repositories for storing metadata, schemas, taxonomies, code lists, ... - Repositories such as Digitaliser.dk in Denmark, the ESD toolkit standards lists in the UK and the European Union repository SEMIC.EU.
#18	2014	Greece + comparison of Point of Single Contact (PSC) of Malta, Slovak	Legal entities and tourism	Use case and comparison to other PSCs.	1. Data and ontology review 2. Ontology enrichment 3. Use case mapping on ERMIS 4. Possible adaptation check for other countries	e-GIF, ERMIS portal, other EU member states tourism PSCs	- Enriched e-GIF ontology for Greek portal ERMIS - Comparison to some other Points of Single Contacts	- Review, matching and enrichment of an ontology

#	Year	Focus region	Public service	Scope (size/type of study/use case or general)	Methodology (list of methods used)	Data (sources)	Outcomes	Relevance to De4a Semantics
		Republic, Spain and Cyprus						
#38	2015	General/EU	Provisioning of geospatial data	Use case on a Greek Spatial Data Infrastructure (SDI)	1. Transformation of INSPIRE metadata to RDF 2. Transformation of INSPIRE data to RDF 3. URIs assignment, specifications, data alignment 4. Exposing data via GeoSPARQL endpoints	Greek SDI's data, INSPIRE metadata, GeoSPARQL specification	Methodology to provide INSPIRE data in a semantic manner via SPARQL endpoints	- Transformation of existing data sources with semantic annotations
#56	2016	Pakistan	Public Procurement Services	Use case of proposed framework evaluation on a Public Procurement Management System	1. Review of Pakistan's government agencies and selection of a use case system 2. Implementation of E-Participation Maturity Model 3. Design of a readiness assessment toolkit No clear evaluation was provided.	Published research + Procurement system by the Sindh Public Procurement Regulatory Authority (SPPRA) for the use case	Authors presented the toolkit which does not seem to have strong theoretical or practical grounds. The toolkit is also not evaluated or discussed to an extent to validate it for the selected use case. Therefore cannot draw strong conclusions based on this paper.	- E-Governance Readiness Assessment Toolkit focused on checking of violation of government policies, rules, and regulations

#	Year	Focus region	Public service	Scope (size/type of study/use case or general)	Methodology (list of methods used)	Data (sources)	Outcomes	Relevance to De4a Semantics
#26	2018	General/ World	All public administration's API-based services	Public services in general	Systematic literature review (they also list/target specific conferences and journals)	Literature (final 26 selected papers for review)	Inputs that are to be used by the Brazilian Government for public services provisioning - four RQs (best practices for digitisation, automation; technological point of view; promotion)	- Review of existing systems, related to DE4A platform and services
#46	2018	General/EU	Linguistic resources access from various languages/countries	European linguistic resources - theoretical idea	Presentation of an example of how to make resources interoperable using existing schemas and how to match data.	EuroVoc, SKOS, LEMON, ONTOLEX schemas	Paper introduces an EU that just started	- Interoperability of specific resources at the EU level (i.e., language resources) - Idea to implement an infrastructure for alignment and interoperability of lexicons
#9	2019	General/ Worldwide Presented use case in Taiwan	General - focus on data exchange Multiple ministries and low-level governments in Taiwan	Study and use case	1. Literature review and hypotheses definition 2. Use case system review in Taiwan 3. Survey of stakeholders in Taiwan 4. Support to hypotheses and models	Different studies/papers	Models for effectiveness, efficiency and accountability; survey results. Usable to check when designing cross-X eGovernment systems.	- Evaluation of performance of cross-border e-government systems

#	Year	Focus region	Public service	Scope (size/type of study/use case or general)	Methodology (list of methods used)	Data (sources)	Outcomes	Relevance to De4a Semantics
#51	2019	General	General	Exploratory literature review since 2006. Selection of final 78 papers.	1. Searching and selection of final 78 papers from various sources. 2. Definition of AI term, techniques. 3. Review of usage of AI in different fields in governments or public sector. 4. Discussion of added values and possible drawbacks.	Literature (78 papers from journals, conferences, books, reports and Web pages).	- Selected definitions of AI. - Literature review since 2006. - Overall discussion pointing out also possible challenges, such as addressing bias in AI, relocation of employees.	- Identification of added values and drawbacks of using AI. - Review of concrete applications in public policy making along with current trends.
#20	2020	General/ World	All public administration's API-based services	Review of usage and challenges of AI in public administration, focus on decision making, chatbots, public governance and security	1. Review of deploying AI for selected focus areas 2. Review of deploying AI with respect to selected challenges 3. Discussion on regulation and governance of AI	Literature/papers	- Review of AI-related challenges and issues - Pointed out focuses on technological and governance innovations that need to be addressed when using AI -> no real clear author's-based views or critical comparison/evaluation was done in the work	- A shallower review of AI usage possibilities and challenges for public administrations

#	Year	Focus region	Public service	Scope (size/type of study/use case or general)	Methodology (list of methods used)	Data (sources)	Outcomes	Relevance to De4a Semantics
#28	2020	General/ World	All public administration's API-based services	Historical development of technological transformation of Government services.	Review of digital governments transformation - from eGovernment to Digital Government.	Journal papers from several databases	Confirming that barriers for transformation are complex and often not technology-related. Using emerging technologies should not be considered. Aspects and effects of transformations are presented.	- Review of important aspects for eGovernment transformation.
#5	2021	Whole EU, Greece, Austria	Legal data processing and merging into one schema and database	All legal documents from EURLex, Austria and Greece	1. Data pre-processing 2. Data analysis with semantic annotation 3. Data translation 4. Data storage	- Greek legal information system - Austrian legal information system - Hellenic Parliament portal - EUR-Lex	An open and automated legal system capable of providing any EU country's legal information based on the existing ontologies.	- Cross-border interaction and information sharing - Aligning/merging schemas of two legal ontologies
#7	2021	USA, Australia, Great Britain, Pakistan and Russia	General - focus on integration of digital and artificial intelligence technologies in public administration	Existing literature for management systems in corporate sector and public administration	Comparing evolutionary stages in adding analytic approaches to the technologies in corporate sector and public administration.	Reports of management systems usage.	Specifics of public administration, its basic differences from governance in the private sector, lead to deviations from the path of digital evolution that is observed in the	DE4A also expects adding analytics (and semantics) and other capabilities to existing technologies in public administration.

#	Year	Focus region	Public service	Scope (size/type of study/use case or general)	Methodology (list of methods used)	Data (sources)	Outcomes	Relevance to De4a Semantics
							private sector.	
#15	2021	General	Social security domain	Schema development (i.e., encoding into semantic representation) for a specific use case	1. Review of data formats 2. Semantic annotation of official forms 3. Extension/encoding of ESSIM in an RDF 4. Use case of specific examples	Existing EU official forms + ESSIM + semantic schemas	Extension of Dublin Core metadata specification for social security data exchange - Exchange Social Security Information Metadata (ESSIM)	- Data exchange and semantic interoperability in social security domain - Usage of semantics and modelling new schemas - Goal of inclusion semantics and specification to information systems development guidelines - Implementation of existing schemas with RDF
#21	2021	General/ World - OECD countries	Tax public administrations	Review of usage of AI and ML for tax processes	1. Review of relevant literature (stages of using of AI in the administration) from different perspectives (user/industry/public sector). 2. Definition of use cases. 3. Proposal of processes and digital tools for tax value chain management.	Literature/OECD data	- Review of a level of AI/ML usage for tax administrations (around 10% of them have automated value chain)	- Extraction of data pipelines/processes for automation

#	Year	Focus region	Public service	Scope (size/type of study/use case or general)	Methodology (list of methods used)	Data (sources)	Outcomes	Relevance to De4a Semantics
#49	2021	Spain	General	Exploratory review/expert interviews	1. Review of theoretical backgrounds for using AI, policies on AI, and AI usage in Spanish administration 2. Selection of AI experts from universities, public administration 3. Interviews 4. Analysis	/	- Definition of risk of using AI and how to address them using law mechanisms. - The need for legal definition of AI.	- Risks and possibilities of using AI in public administration
#19	2022	Brazil	Three main Brazilian public services - ComprasNet (Ministry of Economy), Sisu (Ministry of Education) and Naturalizar-se (Ministry of Justice and Public Security)	Use case for Economy, Education and Justice with generalization to an arbitrary scope	1. Exploratory literature review 2. Interviews with the professors in the field 3. Review of results with the management of selected Brazilian systems	Literature/papers, Polls, Interviews, Information systems management	A use case how to evaluate eGovernment services, along with results for the three selected Brazilian systems	- Evaluation of transparency and accountability of government services
#23	2022	Greece with focus on general EU applicability	All public administration's API-based services	General e-Government interoperable services	1. Literature review 2. Definition of 6-level assessment criteria (legislations and compliances) 3. Update of the schema in two phases (inception - according to the assessment	Declarations, schemas, legislation	Recommendations for assessment a new ontology based on legal, organisational, semantic and technical view)	- Cross-domain, Cross-border API data interchange - Update of a Greek NIF with EIF, OOP and others

#	Year	Focus region	Public service	Scope (size/type of study/use case or general)	Methodology (list of methods used)	Data (sources)	Outcomes	Relevance to De4a Semantics
					criteria + implementation) 4. Definition of assessment methodology			
#30	2022	General/EU, Greek use case	General, use case on procurement	Public services in general, use case presented on one system/public organization	1. Review on the evaluation of assessment models for digital transformation (frameworks, assessment tools, maturity tools). 2. Definition of a model (MOIA), compliant with EIF 3. Validation of the model on Greek Electronic Public Procurement System (ESIDIS)	Legislation, existing literature, ESIDIS	New model for assessment of digital transformation	- Model for assessment that can be used to check the DE4A platform
#53	2022	General/EU	General	Any usage of AI in public administration within 30 EU countries (27 EU member states)	1. Review of different aspects of where AI can be used 2. Review of usages of AI in public administration	250 cases of AI usage across EU found within review articles	- Different types of AI technologies are used for different governance functions (e.g., policy making, service delivery, ...)	- AI improves public service delivery and assist internal organisation management - Interesting review of where, how and which AI technology is used in different public administrations across EU

2.4.2 Demographic distribution and data considered

Most of the selected papers focus on general adoption of machine learning and artificial intelligence in various fields of public administration (11 papers). Five of them especially focus on API-based communication and data exchange. Some of the studies focus on specific use cases, such as legal/justice data processing and schema design (3 studies), procurement services (2 studies), geospatial data provisioning (1 study), social security services (1 study), tax processes (1 study).

Most of the studies are reviews of existing works (13 studies), while five studies present specific use cases in the public services domain. Two studies present specific products - building a semantic information system for eProcurement, and development of a semantic schema along with semantic encodings representations for social security domain.

Our systematic literature review was focused on finding relevant publications since 2012. In the Figure 4, we can see that we extracted one relevant study per year until 2016. More of the selected studies were published in the last two years. As the SLR was performed in mid-2022, we can expect more relevant papers published until the end of this year. Before applying the exclusion criteria also some older papers were found but it seems that the topics of using advanced technologies such as artificial intelligence and semantic Web is gaining importance.

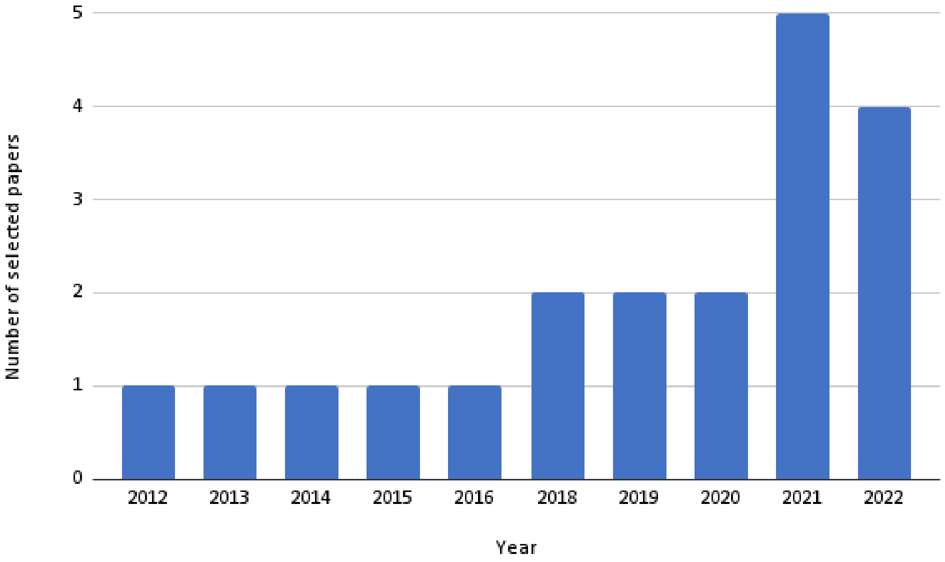


Figure 4: Distribution of selected papers by published year

Studies focused on specific use cases that were developed in specific countries or reviewed existing work in general. Based on the region the studies focused, we can divide them into two high-level clusters:

- European-focused countries (11 studies)
    - Three studies focused on specific European countries:
        - Greece, Malta, Slovak Republic, Spain and Cyprus (1 study)
        - Spain (1 study)
        - Brazil (1 study)
    - Four studies focused on the applicability to the whole EU, some of them with specific focus with use cases from Greece (3 studies) and Austria (1 study).
    - Four studies mainly focused on European applicability with general aspects for implementation of the technologies anywhere in the World.
- Non-European-focused countries or worldwide (10 studies)

- Two studies focused on specific countries:
  - USA, Australia, Great Britain, Pakistan and Russia (1 study)
  - Pakistan (1 study)
- Three studies focused on specific countries use cases with generalisation to the worldwide implementation. One study focused on Taiwan, while other two on OECD countries.
- Five studies were focusing on implementation of technologies in public administration in general with worldwide applicability.

2.4.3 Methodological approaches

The selected papers were published in scientific venues (i.e., journals, conferences, books) and they follow standard structure. Also, the focus of methodologies is to provide justifiable insights that can be drawn based on a selected scientific method. From all the selected studies, we believe that only one of them (study #56) did not provide either transparent evaluations or undoubtedly scientifically proven/supported outcomes.

Methodologies of each specific study is outlined in Table 9. Based on the type of studies we can roughly divide their methodologies in two groups:

- Studies that deal with a specific use case, propose or implement a specific information system that can be used within a public administration. They follow the following structure:
  a. Selection of specific problem or review of a system. Definition of specifications to investigate.
  b. Definition of proposed system architecture, prototype or review of a system that is already deployed in a production environment.
  c. Implementation of the proposed system to at least a prototype version. Gathering feedback from a system (based on logs, metrices) or users (conducting surveys, interviews).
  d. Evaluation with discussion and identification of possible improvements in the future.
- Studies that survey existing work and draw conclusions based on them. They follow the following structure:
  a. Definition of literature review search criteria, studies retrieval.
  b. Specialization of literature review, drawing initial findings.
  c. Analysis and generalization of gathered protocols/models/guidelines.
  d. Discussion and validation of outcomes. The outcomes mostly represent usage scenarios, support to hypotheses, definition of improved models or suggestions for regulation and governance.

2.4.4 Relevance of state-of-the-art for the DE4A semantics

All the studies were published in peer-reviewed scientific venues and therefore represent state-of-the art of applying artificial intelligence and semantics to public administrations and their services. As we searched for multiple criteria and selected most relevant papers via multiple filtering, there exist different aspects of how the studies are related to the DE4A semantics. From the selected 20 studies we can extract the following groups of papers that are related to the project:

- ***Assessment tools and models of advanced technologies adoption***: A group of studies focus on different aspects of measuring levels of adoption of advanced technologies. These include readiness assessment toolkits and models that could be used to evaluate the DE4A pilots. The evaluations are mainly related to (a) checking possible violation of government policies, rules and regulations, (b) risk and possibilities of using AI, and (c) evaluation of transparency and accountability of government services.

- ***Review of digital transformation tools and adoption of AI in public administration information systems***: The second group of studies focuses on existing implementations of AI tools into information systems or deployment of advanced information systems that enable digital empowerment of public services. The studies point out (a) added values and possible drawbacks in using AI-enabled systems, (b) important aspects of eGovernment digital transformation, (c) importance of ETL systems and pipelines. The major outcomes of studies include reviews of existing systems, trends and use cases in specific areas of public administration.
- ***Data schemas management and cross (-domain and -border) data exchange***: This group of studies focuses on interoperability of different resources among institutions and definition of schemas. More specifically studies present (a) an infrastructure for alignment and interoperability of lexicons, (b) processes of existing schemas adaptation to enable data interchange, (c) possibilities to empower existing systems with semantic technologies, and (d) specific use cases show how semantic interoperability enabled easier digitalization of public services.

The last group of papers that we list above (i.e., *Data schemas management and cross (-domain and -border) data exchange*) is the most relevant for the DE4A semantics and therefore we review them more thoroughly. All the studies are briefly presented below.

In Delgado et al. (2021, #15) the authors were researching the issues of international eGovernment data exchanges. They selected the case of social security data exchanges for which they implemented a new metadata specification based on Dublin Core elements. The specification supports international social security exchanges, named Exchange Social Security Information Metadata. Their proposal is based on Linked Data using RDF(S), SPARQL, Microdata and JSON-LD and is planned to be included as a part of an international standard. In the study authors showed an end-to-end process in custom schema development and provided guidelines for developing semantic information systems.

Schmitz et al. (2018, #46) presented the Public Multilingual Knowledge Management Infrastructure action launched by the EC to promote the Digital Single Market in the EU. The project aims to share maintainable and sustainable language resources, and making these resources interoperable in order to support language technology industry, and public administrations, with multilingual tools able to improve cross border accessibility of digital services. The authors conducted a comparative study among the main data models for lexicon representations. They identified a set of tools and facilities (also possibilities of using machine translation technologies) to establish semantic interoperability among multilingual lexicons.

Avgerinos et al. (2021, #5) focused on cross-border data interoperability in the legal domain. They compared legal information systems in Greece, Austria, Hellenic Parliament portal, and EUR-Lex. Current state shows that each country's legal information is currently fragmented across multiple national databases. As it has been shown that design of interoperable legal systems contributes to new advancements, authors proposed an open and automated legal system capable of providing any EU country's legal information based on the currently already existing ontologies.

Kalogirou et al. (2022, #23) researched interoperability as an ability of a product or system to connect with other products or systems without restrictions. Based on the EIF, SDG regulation, and the OOP, the authors update the Greek NIF. The results and the proposed assessment methodology can be reused in other countries and can be further adapted for updating the EIF. The new ontology was created with respect to the legal, organisational, semantic and technical view.

Chen et al. (2019, #9) examined the performance of a cross-boundary e-government system. They reviewed relevant literature and developed a conceptual assessment framework. The key performance measures include efficiency, effectiveness, and accountability. During the empirical evaluation they found out that citizen-centric approach and innovations enhance efficiency and accountability, while administrative interdependence impacts effectiveness and accountability.

Blinova et al. (2021, #7) researched impacts of integrating digital and artificial intelligence technologies in public administration. They compared public sector with corporate sector and revealed similarities that are determined by the key characteristics of modern social processes. The main difference between them is not the lag in public administration (as one could think) but the transparency which violates the various levels of autonomy of systems. The conclusion was that the integration of artificial intelligence in public administration is a key condition for increasing efficiency and ensuring sustainability in the coming years.

2.4.5 Concluding remarks of literature survey

The presented SLR is actually a meta survey as majority of the selected papers were reviews of other works. We hoped that we could find more use case papers or in-depth descriptions or evaluation of existing systems, deployed in public administrations. We believe that many countries already have sophisticated information systems to run their digital transformations but not many of them are described in the scientific literature or probably elsewhere. Based on Digital Economy and Society Index (DESI) we claim that digital transformation across EU countries is happening but obviously each country is probably transforming separately from others. Our conducted SLR reflects that there is a strong need to digitise public administrations using advanced technologies such as tools based on artificial intelligence, machine learning, Semantic Web to provide "*next-generation public services*". As countries already have some of the advanced technologies deployed, rules for data interoperability needs to be set. Also, countries may share their solutions and knowledge by providing reusable building blocks which would speed up the convergence even faster.

Around 20 years ago conferences in the field of computer networks were pushing towards "*converged networks*." These kinds of networks (based on common TCP/IP stack) enabled us to use the same medium, the same underlying technology to make phone calls, watch TV, listen to the radio, ... We believe that the use of Semantic Web technologies is similarly needed to be used by different public administrations to enable "*converged public administrations*" across the EU.

3 Data in the public services

The DE4A system should be monitored and analysed to inform the public administration and users about its trends and performance by means of an analytic component.

Data protection prevents the use of evidence data and other personal data for the purpose of data analytics. However, the DE4A system provides non-personal data with the Logging subsystem (LOG), described in deliverable D5.3, and the Information Desk component (IDK), described in deliverable D3.6.

Considering the available LOG and IDK data, this section proposes a set of metrics to build a proper dashboard for trends and performance of the DE4A system, which are defined semantically for their better understanding and for enabling the automatization of the data extraction, transformation and loading processes besides their analysis.

3.1 Data sources for analytics

Data relevant for the DE4A analytics from DE4A LOG and DE4A IDK are described below.

3.1.1 IDK components as a data source

The DE4A IDK contains available provisions, so there is a biunivocal relation between a provision and a data service. A provision corresponds to a canonical object type -canonical evidence type or canonical event catalogue- provided by a data owner. Besides, the IDK registers some metadata for each data owner, in particular the corresponding Member State and administration level.

3.1.2 LOG component as a data source

The DE4A LOG defines an internal log entry with the syntaxis:

[Timestamp] [Level] [Code] [Logging Participant] [Specific text]

- **Timestamp**: YYYY-MM-dd**T**HH:mm:ss.SSS**Z** (UTC Zulú time)
- **Level**: INFO, ERROR (only levels relevant for the DE4A data analytics)
- **Code**: unique code assigned to the log message type
- **Logging Participant**: ID of the DE4A participant running the logging component
- **Specific text**: text of the message usually corresponding to a message template associated to the code with proper values for the arguments that are included in the template surrounded by brackets.

In the case of messages logged by Data Evaluator and Data Owner components, which code starts with “DE” and “DO” respectively, the specific text is preceded by “[UC#n.m]”, where “n.m” is the code of the DE4A use case run by such components at the time of the logging.

Besides, [Code] is unique by the following syntax, and as elaborated in Table 10:

[Component Code][Level Code][Template 2-digits Number]

Table 10: Component codes and Logs

Component code	Logging component
DR	Data Requestor
DT	Data Transferor
AA	SSI Authority Agent

DE	Data Evaluator
DPO	Data Owner

Level code	Level
I	INFO
E	ERROR
W	WARNING

3.2 Semantics for metrics

Analytical data and metrics should be described with metadata to provide semantics for understanding their role, relevance and meaning, which also contributes to their proper evolution according to the changes and new needs.

There are two main concepts for describing a metric in the context of the DE4A system: raw data describing the dataset of the metric, and dashboard indicators created from the metric's raw data. The semantics for describing metrics through their raw data and dashboard indicators can use a formal language to be processed by machines, which is not covered in this deliverable.

Besides, raw data may be used to, for instance, predict demand peaks and detect component malfunctioning by means of advanced algorithms. Currently, because of the few logs generated by the DE4A pilots, it is not possible to develop this feature.

3.2.1 Raw data

The "Raw data" of a metric is the resulting dataset from specific data sources after some extraction, transformation and loading processing. There is just one raw data per metric so the DE4A analytical component stores all the datasets collected according to each raw data defined.

Each dataset is composed by a set of parameters and one or more basic measures calculated as a function of all the parameters. In consequence, the resulting dataset can be used to create a dynamic table to analyse the data by their combination and filtering.

Following, the metadata to describe the "Raw data" of a metric is explained:

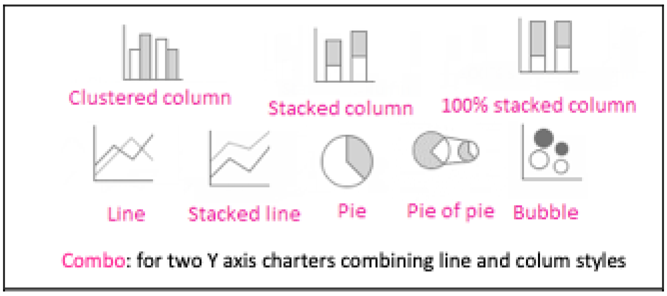
- **Code**: "M{n}.DTBL", where {n} is a sequential number so M{n} is the code of the corresponding metric.
- **Name**: meaningful title of the dataset
- **Purpose**: description of the purpose of the dataset in the context of the DE4A analytics
- **Data source**: source of the data to collect
- **Frequency**: frequency of the data collection
- **Param**: category of parameters of the dataset. The dataset can have more than one category of parameters, each one sequentially numbered.
- **Subparam**: parameter within a specific category. The dataset can have more than one sub-parameter per category, each one sequentially numbered within the category.
- **Measure**: qualitative or quantitative function considering all the subparams. The dataset can have more than one measure, each one sequentially numbered. In particular, the measure "count of occurrences" is the number of unique combinations of subparam values in the dataset, while "sum of occurrences" is the number of combinations of subparam values including identical combinations.

3.2.2 Dashboard indicator

In this context, the “Dashboard indicator” represents a chart that shows specific data from the raw data of a metric. Dashboard indicators can provide pieces of relevant information in an easier way for humans to process it at a glance.

In this context, a Dashboard indicator is described by the next metadata:

- **Code**: M{n}.D{m}, where M{n} is the code of the corresponding metric and {m} is a sequential number with M{n}.
- **Name**: meaningful title of the information represented.
- **Description**: description of the information represented.
- **Style:** style of the chart. For instance, observe Figure 5.



Combo: for two Y axis charters combining line and colum styles

Figure 5: Chart styles for visualisations

- **Axis X**: horizontal axis of the chart. If more than one, X{n} corresponds to the parameter whose values are shown per each X{n+1} parameter value.
- **Axis Y**: vertical axis of the chart. For Combo style charts, axis “YC” corresponds to the column values and axis “YL” corresponds to the line values.
- **Series**: for pie charts and charts with more than one line or column, each colour represents a value of the sub-parameter(s) used for the series of the chart.
- **Value**: function of the measure for the corresponding combination of parameters. For combo charts, “VC” is for column values and a “VL” is for line values.
- **Cumulative:** if the value is represented in a cumulative way through a time dimension.
- **Target**: maximum or minimum threshold or target for the value. When a Dashboard indicator has a target, it represents a Key Performance Indicator (KPI) of the system.

3.2.3 DE4A trend and performance metrices

Following, the proposed metrics to monitor the trend and performance of the DE4A System is described according to the semantics defined above. The metrics are aimed to provide information on:

1. Cross-border provision onboarding
2. Exchange transactions according to the different interaction patterns
3. Exchange process errors
4. Business events

3.2.4 M2: Metric on the cross-border provision onboarding

The cross-border provision onboarding process has raw data as described in Table 11.

Table 11: Raw Data of cross-border provision onboarding

Code	M1.DTBL	
Name	**Notified cross-border provisions**	
Purpose	Trimestral information on the **onboarding process regarding both cross-border services and competent authorities**. The provision onboarding process cannot set thresholds or target values because its only obligation is to make cross-border available existing national services that are equivalent to canonical types. In this regard, metric M1 can only be informative.	
Data Source	IDK provisions	
Frequency	Data collected the first day of each trimester	
Param 1	PROVIDER	Competent authority that provides the service
Subparam 1.1	- Preferred label	Preferred label of the provider
Subparam 1.2	- Member State	Member State of the provider
Subparam 1.3	- Administrative Level	Administrative level (national, regional, local, educational) of the provider's competences
Param 2	CANONICAL OBJECT TYPE	Type of canonical object provided by the service
Subparam 2.1	- Category	EVENT, for canonical event catalogues EVIDENCE, for canonical evidence types
Subparam 2.2	- Canonical name	Short name of the canonical object type
Subparam 2.3	- URI	Complete URI of the canonical object type. An ID token may correspond to more than one URI if there are several versions of such canonical object type.
Param 3	TIME OF MEASURE	Moment of the data collection
Subparam 3.1	- Year	Year of the day before the data collection
Subparam 3.2	- Trimester	Trimester of the day before the data collection
Measure 1	Count of occurrences	The dataset cannot include identical combinations of values of Param 1 and Param 2 sub-parameters

Based on the data of Table 11, we can create dashboards with indicators as illustrated in Table 12.

Table 12: Dashboard indicators: cross-border provision onboarding

Code	M1.D1	M1.D2
Name	**Evolution of notified providers**	**Evolution of notified services**
Description	Cumulative yearly notified cross-border providers per administrative level and member state	Cumulative yearly notified services by their provider member state and admin level
Style	Stacked column	Stacked column
Axis X1	Year (3.1)	Year (3.1)
Axis X2	Member State (1.2)	Member State (1.2)
Axis Y	Nº Providers	Nº Services
Series	Admin. Level (1.3)	Admin. Level (1.3)
Value	Sum of occurrences	Sum of occurrences
Cumulative	Yes	Yes
Target	None	None

3.2.5 M2: Metric on exchange transactions

The information exchange process residues log data in the systems it uses for exchange. Table 13 describes the nature and types of these data.

Table 13: Raw data: Metric on exchange transactions

Code	M2.DTBL
Name	**Exchange transactions**
Purpose	Monthly information on evidence exchanges between data evaluators and owners performed under Intermediation (IM), User-Supported Intermediation (USI), Lookup (LU) and Verifiable Credential (VC) patterns, as well as transactions under Subscription and Notification (SN) pattern, along with their participants, canonical object types and success.
Data Source	LOG entries with level "INFO". For Intermediation patterns and subscription pattern Logs entries generated by Data Requestor components when receiving request from Data Evaluators identified by {RequestID} signal the starting of an exchange. Logs entries generated by Data Requestor components when receiving request from Data Owners when receiving a response with the same {RequestID} signal the ending of such exchange. Since each exchange may include several items to request a canonical object type (except in the case of Legacy IM exchanges), an exchange transaction corresponds to a singular request item within a message exchange between a Data Evaluator and a Data Owner.

Exchange started	Exchange ended
[DRI01] Legacy IM Request message received	[DRI06] Legacy IM Response message received
[DRI02] IM Request message received	[DRI07] Evidence Response message
[DRI03] USI Request message received	
[DRI05] LU Request message received	
[DRI04] Subscription Request message received	[DR09] Subscription Response message received
[DRI10] Event Notification message received	n/a

The text templates all of these log entries share the next same relevant set of arguments:

Template Argument	Description
{RequestID}	UUID for the evidence exchange request included in the IEM messages
{DataEvaluator (ParticipantID)}	URI (i.e., participant ID) of the data evaluator for the evidence exchange request
{DataOwner (ParticipantD)}	URI (i.e., participant ID) of the data owner for the evidence exchange request
({list (RequestItemId):(Canonical ObjectTypeUri)}) OR {CanonicalObjectTypeUri}	List of items included in the IEM exchange request message, specifying the item ID and requested canonical object type URI for each of them. Legacy IM messages do not support multi-item requests, so there is only one canonical evidence type. In the case of log signalling the end of the exchange, an error code may substitute the canonical object type URI. A canonical object type is a canonical event catalogue for exchanges under the SN pattern; for the rest of the exchanges it is a canonical evidence type.

For Verifiable Credential pattern

In this case there are two kinds of transactions: between the edge agent (the user's wallet) and the data evaluator or the data owner respectively. Log entries generated by Authority Agent components signal the starting and ending of such transactions.

The starting of both kinds of transactions are signalled by the log entry:
[AAI03] Generated DID invitation for edge agent with ID {UUID}
where {UUID} identifies the session within which a verifiable credential is not only presented but also accepted, so this last event signals the ending of an evidence exchange transaction. In this case, each evidence exchange transaction is identified by {VC Id}, which connects log entries that present and accept evidence corresponding canonical evidence type in the form of a verifiable credential.

VC presented	VC accepted
[AAI22] Sent Offer for Verifiable Credential {VC Id} of type {CanonicalEvidenceTypeUri} under invitation {UUID} from {DO URI}.	[AAI06] Sent Verifiable Credential {VC Id} to the edge agent under invitation {UUID} from {DO URI}.
[AAI20] Received Verifiable Credential {VC Id} at the verifier {DE URI} under invitation {UUID}	[AAI21] Validated Verifiable Credential {VC Id} of type {CanonicalEvidenceTypeUri} under invitation {UUID} for {DE URI}.

Details on the data evaluators and owners are obtained from the IDK

Frequency	Data collected the second day of each month selecting messages logged the month before	
Param 1	TRANSACTION (Tx)	Each element of the list argument in the log entry template for IM, USI, LU and SN patterns corresponds to one exchange transaction, and they share the arguments of the log message template other than the ones included under the “list” section. For the Legacy IM and VC patterns, each log entry corresponds to one transaction.
Subparam 1.1	- Pattern	Pattern of the Tx: - IM [Code] in ([DRI01], [DRI02]) - USI [Code] -> [DRI03] - LU [Code] -> [DRI05] - VC [Code] -> [AAI03] - SN [Code] in ([DRI04], [DRI10])
Subparam 1.2	- Subpattern	Subpattern of the transaction if any: - IM Legacy [Code] -> [DRI01] - VC Issuing [Code] -> [AAI22] with argument {UUID} equal to the “Request ID” of the Tx (subparam 1.3) - VC Verifying [Code] -> [AAI20] with argument {UUID} equal to the “Request ID” of the Tx (subparam 1.3) - SN Subscription [Code] -> [DRI04] - SN Notification [Code] -> [DRI10] - Empty otherwise
Subparam 1.3	- Request ID	Argument {RequestID}, {NotificationID} or {UUID} in the log entry used to obtain the “Pattern” of the Tx.
Subparam 1.4	- Item ID	When the Tx has not subpattern or it is “Subscription”, argument {RequestItemID} of one element from the argument “list” that is part of the log entry used to obtain the “Pattern” of the Tx. When “Subpattern” of the Tx is “Notification”, argument {NotificationItemID} or one element from the argument “list” of the log entry used to obtain the “Pattern” of the Tx.

		When “Subpattern” of the Tx is “Legacy”, the same value than “Request ID” (1.3). When “Pattern” of the Tx is “VC”, argument {VC Id} in the log entry used to obtain the “Subpattern” of the Tx.
Param 2	**CANONICAL OBJECT TYPE**	
Subparam 2.1	- Canonical URI	Argument {CanonicalEvidenceTypeUri} in the log entry used to obtain the “Subpattern” (1.2) of the Tx when it is “IM Legacy” or “VC Issuing”. When “Subpattern” is “VC Verifying”, such an argument is specified in the log entry of code [AAI21] whose argument {UUID} is the “Request ID” (1.3) of the Tx. Otherwise, the argument {CanonicalEventCatlogueUri} or {CanonicalEvidenceTypeUri} that is the paired with the “Item ID” (1.4) of the Tx as an element of the argument {list ()} in the log entry used to obtain the “Pattern” (1.1) of the Tx.
Subparam 2.2	- Canonical Name	Short name of the canonical object type extracted from “Canonical URI” (2.1)
Subparam 2.3	- Canonical Version	Short name of the canonical object type extracted from “Canonical URI” (2.1)
Param 3	DATA EVALUATOR	End participant that requests the exchange Tx.
Subparam 3.1	- Data Evaluator URI	Argument {data evaluator (participant id)} or {URI DE} in the log entry used to obtain the “Pattern” (1.1) or “Subpattern” (1.2) of the Tx.
Subparam 3.2	- Evaluator Country	Country associated to the “Data Evaluator URI” (3.1) in the IDK.
Subparam 3.3	- Evaluator Admin Level	Administration Level associated to the “Data Evaluator URI” (3.1) in the IDK.
Param 4	DATA OWNER	End participant that provides the requested element.
Subparam 4.1	- Data Owner URI	Argument {data owner (participant id)} or {URI DO} in the log entry used to obtain the “Pattern” (1.1) or “Subpattern” (1.2) of the Tx.
Subparam 4.2	- Owner Country	Country associated to the “Data Owner URI” (4.1) in the IDK.
Subparam 4.3	- Owner Admin Level	Administration Level associated to the “Data Owner URI” (4.1) in the IDK.

Param 5	TIME	
Subparam 5.1	- Starting Timestamp	[Timestamp] of the log entry used to obtain the “Pattern” (1.1) of the Tx.
Subparam 5.2	- Year	Year derived from “Starting Timestamp” (5.1)
Subparam 5.3	- Month	Month derived from” Starting Timestamp” (5.1)
Subparam 5.4	- Trimester	Trimester derived from “Month” (5.3)
Param 6	RESPONSE	
Subparam 6.1	- Ending Timestamp	When possible, timestamp of the log entry that signals the end of the Tx as described above: – [AAI06] with arguments {UUDI} and {VC Id} equal to the “Request ID” (1.3) and “Item ID” (1.4) of the Tx. – [AAI21] with arguments {UUDI} and {VC Id} equal to the “Request ID” (1.3) and “Item ID” (1.4) of the Tx. – [DRI06] with argument {RequestId} equal to the “Request ID” (1.3) of the Tx – [DRI07] with argument {RequestId} and {RequestItemId} equal to the “Request ID” (1.3) and “Item ID” (1.4) of the Tx – [DRI09] with argument {RequestId} and {RequestItemId} equal to the “Request ID” (1.3) and “Item ID” (1.4) of the Tx If the Tx is under subpattern (1.2) “Notification”, the value of “Starting timestamp” (Subparam 5.1) Otherwise, empty.
Subparam 6.2	- Error	“COMM” if the “Ending Timestamp” (6.1) of the Tx is empty, since the error is due to some communication problem between the participants. When the “Elapsed time” can be calculated, the argument {ErrorCode} paired with the “Item ID” (1.4) of the Tx in the argument {list ()} of the log entry used for such calculation, if such an argument is present. Otherwise, “NO ERROR”.
Measure 1	Elapsed time	When possible, elapsed time between the “Starting Timestamp” (5.1) and “Ending Timestamp” (6.1) of the Tx.
Measure 2	Sum of occurrences	

Table 14: Dashboard Indicators on significant exchanges

Code	M2.D1	M2.D2
Name	**Evolution of successful requests per requesting member state and pattern**	**Evolution of successful requests per providing member state and pattern**
Description	Cumulative trimestral number of requests and responses per requesting member state	Cumulative trimestral number of requests and responses per providing member state
Style	Combo 100% stacked column	Combo 100% stacked column
Axis X1	Year (5.2) and Trimester (5.4)	Year (5.2) and Trimester (5.4)
Axis X2	Evaluator Country (3.2)	Owner Country (4.2)
Axis YC	Nº Requests	Nº Requests
Axis YL	Nº Responses	Nº Responses
Series	Pattern (1.1) and Subpattern (1.2)	Pattern (1.1) and Subpattern (1.2)
Value C	Sum of occurrences	Sum of occurrences
Value L	Sum of occurrences with subparam 6.2 “NO ERROR”	Sum of occurrences with subparam 6.2 “NO ERROR”
Cumulative	Yes	Yes
Target	None	None

Code	M2.D3	M2.D4	M2.D5
Name	**Evolution of successful exchange transactions per canonical type and pattern**	**Evolution of successful exchange transactions between member states**	**Evolution of unsuccessful exchange transactions between member states**
Description	Trimestral number of requests per canonical object type, pattern and success, combined with the average elapsed time of the response for non-VC patterns	Trimestral number of successful requests between member states	Trimestral number of unsuccessful requests between member states
Style	Combo stacked column	Combo stacked column	Combo stacked column
Axis X1	Pattern (1.1) and Subpattern (1.2)	Evaluator Country (3.2)	Owner Country (4.2)

Axis X2	Canonical name (2.2)	Owner Country (4.2)	Trimester (5.4)
Axis X3	Trimester (5.4)	Trimester (5.4)	Year (5.2)
Axis YC	Nº Exchanges	Nº Exchanges	Nº Tx
Axis YL		Average minutes	
Series	Year (5.2)	Year (5.2)	Error (6.2)
Value C	Sum of occurrences with subparam 6.2 “NO ERROR”	Sum of occurrences with subparam 6.2 “NO ERROR”	Sum of occurrences with subparam 6.2 not equal to “NO ERROR”
Value L		Elapsed time in minutes	
Cumulative	No	No	No
Target	None	None	None

3.2.6 M3: Metric on the exchange process errors

In DE4A, errors in the process are looed using a specific format as described in Table 15.

Table 15: Raw data - exchange process errors

Code	M3.DTBL	
Name	**Exchange process errors**	
Purpose	Monthly information on errors logged by a DE4A Log component.	
Data Source	LOG messages with [Level] = [ERROR].	
Frequency	Data collected the second day of each month selecting messages logged the month before	
Param 1	ERROR TYPE	
Subparam 1.1	- Code	Field [Code]of log entry of level “ERROR”.
Subparam 1.2	- Message	Field [Specific text message] of the log entry used to obtain the “Code” (1.1) of the error.
Subparam 1.3	- Logging participant	Field [Logging participant ID] of the log entry used to obtain the “Code” (1.1) of the error.
Subparam 1.4	- Logging component	Corresponding to the first two characters of the “Code” (1.1) of the error: *DR: Data Requestor* *DT: Data Transferor* *AA: SSI Authority Agent* *DE: Data Evaluator*

		DO: Data Owner
Param 2	ERROR TIME	Timestamp of the error
Subparam 2.1	- Year	Year extracted from the field [Timestamp] of the log entry used to obtain the “Code” (1.1) of the error.
Subparam 2.2	- Month	Month extracted from the field [Timestamp] of the log entry used to obtain the “Code” (1.1) of the error.
Subparam 2.3	- Day	Day extracted from the field [Timestamp] of the log entry used to obtain the “Code” (1.1) of the error.
Subparam 2.4	- Hour	Hour extracted from the field [Timestamp] of the log entry used to obtain the “Code” (1.1) of the error.
Subparam 2.5	- Day of week	Weekday corresponding to the “Month” (2.2) and “Day” (2.3) of the error.
Subparam 2.6	- Hour Range	“08:00-14:00”, “14:00-20:00“, “20:00-02:00”, or “02:00-07:59” if subparam 2.4 is equal or greater than the beginning of one range and less than the end of such a range.
Measure	Sum of occurrences	

The visualisations can be developed based on the raw data as described in Table 16.

Table 16: Dashboard indicators exchange process errors

Code	M3.D1	M3.D2
Name	**Evolution of errors by category**	**Evolution of time distribution of errors**
Description	Monthly number of errors by participant	Monthly evolution of distribution of errors considering day of week and hour range
Style	Stacked column	Stacked column
Axis X1	Year (2.1)	Day of week (2.5)
Axis X2	Participant (1.3)	Hour Range (2.6)
Axis X3	Component (1.4)	Year (2.1)
Axis Y	Nº errors	Nº errors
Series	Month (2.2)	Month (2.2)
Value	Sum of occurrences	Sum of occurrences
Cumulative	No	No
Target	None	0

3.2.7 M4: Metric on Business events

This metric will summarise the data about the business events as described in Table 17.

Table 17: Raw data Business events

Code	M4.DTBL	
Name	**Business events**	
Purpose	Information on business events classified by the log system: requested evidence is not available yet (but it will), the preview has been rejected by the user at the DO or at the DE side, and a DE4A pilot process has started or ended.	
Data Source	LOG of level “INFO” or “WARN” with a code starting by “DO” or “DE”.	
Frequency	Data collected the second day of each month selecting messages logged the month before	
Param 1	EVENT	
Subparam 1.1	- Code	· *[DOW02] evidence not available yet* · *[DOW06] preview at DO side rejected by the user* · *[DEW04] preview at DE side rejected by the user* · *[DEI01] DE4A pilot process started*
Subparam 1.2	- Message	
Param 2	EVENT DATE	
Subparam 2.1	- Year	Year of log message
Subparam 2.2	- Month	Month of log message
Subparam 2.3	- Day	Day of log message
Measure	Sum of occurrences	

Based on these data we can propose some dashboard indicators for business events as presented in Table 18.

Table 18: Dashboard Indicators Business events

Code	M4.D1	M4.D2
Name	Evolution of DE4A pilot launches	Evolution of DE4A preview rejected by user
Description	Cumulative monthly evolution of DE4A pilot process started combined with the number of “evidence not available yet” occurrences	Evolution of DE4A preview rejected by user
Style	Combo clustered column	Stacked column

Axis X1	Month (2.2)	Month (2.2)
Axis X2	Year (2.1)	Year (2.1)
Axis YC	Nº pilot launches	Nº Preview Rejections
Axis YL	Nº evidence responses delayed	
Series		Message (1.2)
Value C	Sum of occurrences with Code (1.1) equals to “[DEI01]”	Sum of occurrences with Code (1.1) in ([DOW06], “[DEI01]”)
Value L	Sum of occurrences with Code (1.1) equals to “[DOW02]”	Sum of occurrences
Cumulative	Yes	No
Target	None	None

3.3 Conclusion and further work

The DE4A system can be analysed by means of this data analytics proposal, so the trends and performance of the system can be monitored. The proposed dashboard indicators can provide insights on the evolution of the onboarding and the most demanded exchanges, so Member States may take advantage of them in order to develop public policies that boost and ease those exchanges. The proposed dashboard indicators can also provide insight on the most common errors to help participants to focus on the mitigation measures needed.

A formal language may be developed to describe in detail raw data and dashboard indicators. In this way, the extraction, transformation and loading of data from the data sources into the analytical storage can be automated, as well as the periodical generation of the indicator charts to publish in the DE4A dashboard. The DE4A dashboard could be publicly available through a web site.

As new information is needed on the trends and performance of the DE4A system, its log component may be enriched with new log entries that can provide the required data for the new analytical information.

With this proposal, trends and performance of the system can be monitored but AI can also provide relevant information from the Raw data described in this chapter. For example, the storage where the proposed raw data extracted from the mentioned data sources would be loaded may be exploited by predictive algorithms to help participants to properly dimension their infrastructures according to the expected demand. Besides, clustering algorithms may identify specific components and even times where errors are most common.

4 Vision for Machine Learning (ML) practices

This chapter includes a summary of current and potential ML and Data/Text Mining (DM/TM) approaches which would be applicable for enhancing the automated public service across borders. ML applications on top of data and information of public services allow increased efficiency and quality of the services, as well as be able to use for detection of irregularities and anomalies in the technical process. ML also can be used for decision making, load balancing or for other business perspective requirements. Hence, this chapter brings in the knowledge from literature and the information of DE4A data structures and models together to visualise the potential ML applications, in the light of the scarcity of such disruptive technologies for Semantic interoperability.

4.1 Extensions from the SLR outcomes

The systematic literature study (SLR) in Chapter 2.1 clearly showed how little evidence is actually available on the subject of interoperability as a whole and specifically semantic interoperability. Resources in academic literature which scientifically investigates the applications of ML is extremely scarce. As a matter of fact, the SLR focusses on the AI applications in the domain in general. The main outcome from the SLR is that the gap exists in the state-of-the-art AI solutions for enhanced semantic interoperability. We present some solutions that can be useful to increase the harmonisation in the extent of the existing resources in this context.

While AI is quite generally accepted, understood and already deployed within specific scenarios, semantics is lagging behind. Semantic interoperability is mostly based on the Semantic Web technologies which became popular in the 1990s but stalled in 2010s. At a time, large semantic graphs were developed (e.g., DBPedia) and some companies started to manually curate large knowledge graphs (e.g., Google and Freebase). Initial approaches that focused deeply in higher-order semantics, reasoning and linking did not scale well. Industry saw no added value in using semantic approaches and therefore semantic databases and semantic tooling were not developed for production (more about semantic toolkits is presented in D3.6). On the other hand, NoSQL databases, such as graph databases were being developed, along with specialized query languages instead of SPARQL support. Recently, knowledge graphs started gaining popularity and semantic technologies are being revived. Trends can be also observed by large vendors that presented semantic databases in the last two years, for example Amazon Neptune or Oracle Graph Database.

4.2 Extensions from DE4A pilot settings

Public services can be considered as a valuable depository of authority related, infrastructure, operational, statistical, and archival data. The DE4A system has two components that can be considered as data sources: A) LOG and B) IDK component (mentioned in chapter 3).

In this section, we examine the possibilities to apply ML and DM methods to DE4A pilot data in order to extract meaningful insights. Towards this direction a study on existing applications and best practices was conducted. We observed that both industry and academia adopt automated approaches mainly for understanding log data and simplifying troubleshooting.

The existence of a biunivocal relation between an IDK provision and a data service and the relations implied in exchange transactions give the floor for further analysis and transformation of the data aiming at the extraction of valuable information that could benefit the cross-border landscape. Current trends in database manipulation and applications consider ML to better leverage their data and deliver analytical insights.

From sourcing data to making predictions, ML is valuably enhanced using knowledge-graph technology, that includes a collection of interlinked descriptions of concepts, entities, relationships and events that consists of underlying semantics definitions. Considering that knowledge-graph technology is a widely available and a well-understood technology for knowledge representation and reasoning, one possible direction could be the representation of basic concepts and primary metadata in the form of a graph/ontology. Graph algorithms help to identify meaningful graph-oriented metrics and patterns that can be applied widely. These metrics include community detection, closeness centrality, betweenness centrality, and similarity of neighbourhoods. Graph data can also help with clustering as it can detect whether certain nodes form a community. Therefore, making it possible to automatically retrieve analytics information like for example “all the events that belong to a specific provider” or “most frequent communications per country”.

Graph ML could be also used for security purposes to identify patterns of system errors (e.g., times where errors are most common) and relations among fraudulent behaviour and bad actors. Looking for correlated anomalies ensures that there is no coincidental or accidental occurrence of random anomalies in logs.

Another possible approach that could be followed in terms of raw data analysis is pattern recognition based on text similarity methods. A possible idea behind this approach could be the identification and learning of commonalities and differences among the provisions on national level in order to 1) harmonize the main concepts and 2) create a canonical ontology in terms of public administration events. In this case scenario there is also need for Natural Language Processing (NLP) techniques like open information extraction and entity linking that are based on ML models trained on these tasks.

As we have already mentioned, log data has been extensively exploited for system troubleshooting. Log data contains parameters describing the current system state and thus is suitable for driving actionable insights for security reasons. Typical example applications include the detection of anomalous behaviour and system failure.

Anomaly detection: ML can be used to automatically find clusters of anomalies across logs that can be used to automatically detect software problems without any manual training. Applying trained models and AI it is possible to identify systemic and anomalous behaviour patterns. For example, a useful prediction might be to classify whether a particular log event, or set of events, is causing a real incident, such as that requires attention.

System failures detection: Failure analysis allow us to understand system failure modes, establish the cause of failures, prevent their occurrence, and improve the dependability of future system releases. Failure analysis is often conducted by collecting event logs. Event logs report errors that may lead to failure of system and help to explain the root cause of an issue. This could happen by parsing the data. Specifically, determining which parts of the log lines correspond to constant (textual) parts, and which correspond to parameters such as logging IDs, timestamp and Codes for severity (e.g. INFO, ERROR, WARNING).

Machine- Learning approaches

Log data are sequence data that consist of a textual combination of log keys like a time stamp, the identifier of the source of the event, a severity level (e.g., warning or error) and a text message. A typical machine learning pipeline consists of three steps: pre-processing of log keys, feature embeddings to represent a sequence of log keys and lastly, prediction of anomalous behaviour. For the first step, a simple parsing process is conducted to transform log messages into keys. Then a feature extraction approach is used such as Term Frequency–Inverse Document Frequency (TF-IDF) or deep learning method such as Recurrent Neural Network (RNN) models considering log sequences as natural language. Recent methods involve using transformer-based models like Bidirectional Encoder Representations (BERT) which has shown excellent performance in NLP.

Based on the extracted features ML approaches are performed to detect anomalous sequences. Considering the scarcity of anomalous sequences, manually collecting and labelling large amounts of anomalous data is not realistic to train a model and thus there is lack of sufficient data to train a model and follow a supervised classification approach. Therefore, a better reflect approach is to adopt an unsupervised method where a model is trained using only normal log data at the times when there are none or only minimum number of abnormal logs; anomaly data is then used only during verification and testing processes.

Although logging may seem a trivial activity, the analysis and interpretation of log data becomes a complex task. Below we describe some significant challenges associated with applying machine learning to system logs.

Challenges

- The main issue with processing system logs is that log data are unstructured, and their format and semantics can vary significantly from system to system.
- To build a well-performing machine learning model, the data scientist or engineer needs to understand the domain well. This is required to be able to select the correct features, evaluate data sets and understand correlations between data items.
- The use of different types of logging helps to interpret the data from each log. In the case of machine learning, however, spreading relevant system information over multiple system log files significantly complicates the training process. This is because it requires combining pre-processed and analysed data from multiple logs in order to reach reliable conclusions and to infer useful classifications or predictions.
- Processing of the logs with ML techniques in Real-Time (or in foreseeable time) is associated with high computational complexity, especially when the system is dealing with a large number of logs (because ML techniques can be very demanding in terms of resources and time).

4.3 Promising extensions for effective semantic interoperability

4.3.1 Semantic search tool

Semantic search tool assist ontology engineers during ontology development on finding related vocabularies/classes/properties for given concepts.

The collaborative provision of public services among Member States is essential to reduce costs, burdens and barriers for European citizens and businesses, and the implementation of the OOP at European level is the cornerstone for this goal. However, evidence is usually represented by data structures that differ from Member State to Member State, or even some Member States present evidence through electronic documents that prevent machines to process the evidence. Besides, evidence as both data structure and electronic document may not only include different properties, but properties with different meaning depending on the issuing Member State. There are cultural and legal differences between Member States that can mislead the cross-border understanding of some terms. In the case of evidence requested by public administrations, it proves facts as provided by law and in accordance with the applicable legal framework. For instance, a birth has different meaning and characteristics depending on whether it is used in reference to Biology or to Civil Registries, since in the latter case a “birth” is a registration event according to the national law. Same consideration applies for usual terms like “natural person”, “household”, “family”, “income”, “unemployed”, “pensioner”, etc.

Because of the abovementioned issues posed by the cross-border use of evidence, semantics play a central role to enable the required interoperability. Unfortunately, automatic translation tools cannot tackle this challenge since they are aimed to translate natural language, within a given context and without a full precision in the result. These issues are the main reason why many Member States requires the legal translation of evidence issued in a language that is unofficial in the Member State. Some European initiatives have tackled the burden posed by legal translations, such as the multilingual standard forms for public documents under the Regulation 2016/1191 or the use of common data structures in information exchange systems as BRIS. However, the former is a solution that does not include digitalization requirements, and the latter is a solution that requires a significant effort to achieve interoperability agreements and thus is not easily extended to domains with more heterogeneity. Therefore, other semantic assets should be used to define common dataspaces to foster the implementation of the OOP at European level in a realistic and actionable way. There is a need for identifying a canonical data structure that provides evidence equivalent to domestic evidence that are originally issued either as data structures or electronic documents. In this regard, semantic assets with existing knowledge can be reused to define such canonical data structures with a precise and common understanding among Member States.

Recent advances in ML, deep learning, and NLP have shown great capabilities for text processing, text mining and extracting structured information which shows in delivering state-of-the-art results in most of the NLP tasks. A characteristic and disruptive example of such enablement is the BERT architecture that is able to process and understand the syntactic and semantic information in the text. Therefore, it goes one step further than the traditional keyword or ngram-based approaches and is able to detect synonyms and similar phrases but also the deeper meanings in the text. Currently, BERT is the core part in most of the NLP tasks. Furthermore, with the vast number of textual documents in different languages (e.g., EU legislative documents), it is possible to train a language model for predicting words given a context that is able to capture the semantics of the text in multiple languages. In this way, the model is able to understand the meaning of each word based on the context in a sentence. Such model allows the implementation of a semantic search engine where pretrained models for it have already been introduced [15]

Therefore, a machine learning tool may help semantic experts to analyse synergies and reusability of existing semantic assets with the aim to define a common dataspace for cross-border evidence relevant for the collaborative provision of public services at European level. The input of such a tool would be semantic assets as ontologies, core vocabularies, controlled vocabularies, authoritative lists, and taxonomies defined at national or international level. From their explanatory documents, thesaurus and dictionaries automatically translated to English, as a common language, the tool could build a graph of related terms. The graph could be used to locate those terms that better suit to represent a concept introduced by a semantic expert who, at the same time, contributes to improve the graph by selecting the terms that he or she considers appropriate to represent the concept.

To provide additional connectivity to the graph specific thesaurus like Eurovoc can also be employed for classifying evidence descriptions to domain concepts that will further allow the design of the canonical data structure. This can be achieved by applying BERT for multilingual semantic similarity between evidence descriptions and concept labels and definitions or by training a topic classifier using the MultiEurLex dataset that consists of 65.000 EU laws in 23 official EU languages manually annotated with Eurovoc concepts by the EU Publication Office [16]. Apart from supervised learning methods, zero-shot learning techniques have been studied for cross-lingual transfer in legal topic classification where it is indicated that translation-based approaches have shown state-of-the-art results.

Such an ML-based model assumes the existence of rich content in lexical and semantic resources. However, in some cases, linked open vocabularies do not include rich description of the ontology concepts. Nevertheless, most of the linked open vocabularies include a document with the specification of the ontology with information about concept definitions, range, subject, usage notes and more that can be utilised for applying semantic search with BERT. Such documents usually appear in the form of structured HTML (e.g. https://www.w3.org/TR/vocab-dcat-2/) or PDF documents which can be integrated in an ETL (Extract Transform Load) process for extracting metadata and structured information from the content.

4.3.2 Matching information from documents to linked data for an interpretable search engine

With the emergence of the Web and digital transformation, a huge amount of information is stored digitally. However, this information usually exists in the form of textual documents or html pages which is highly unstructured creating an impediment to semantic interoperability. Tim Burners Lee, the inventor of the Web and Linked Data initiator, suggested a 5-star rating for linked data. Textual documents are classified as 2-star data, meaning being "available as machine-readable structured data, (i.e., not a scanned image)", thus being far from 5-star rating of linked data without ensuring high data quality and semantic interoperability. The 5-star rating means that linked open data are "published using open standards from the W3C (RDF and SPARQL)". Therefore, there is a need for a process to transform and extract structured information from documents and match it to linked open data. On top of that, this process needs to be automated as it would require extensive human labour to annotate millions of documents.

To that end, ML, NLP and AI in general, can play an important role for automating this process. As already indicated, the BERT architecture has performed state-of-the-art results in most of the NLP tasks, like information extraction and linking. However, these NLP tasks, that are based on ML, require a huge amount of labelled data, which in many cases does not exists. On top of that, aspects like multilinguality, low-resource languages and different domains of documents (e.g., news articles, legislative documents, tweets, etc.) constitute an impediment for training ML models.

Nevertheless, as there is a huge amount information in textual form, it is possible to train a language model (e.g., BERT) to semantically and syntactically process and interpret the text. Language model is a task for predicting a word given a context. It is considered a semi-supervised approach as there is no human supervision but the training dataset is created automatically by randomly removing words from the text and learning to predict missing words. This technique allowed training large deep learning architectures on huge amount of data. Furthermore, these pretrained models are rapidly used for transfer learning (re-using a pre-trained model on new data). In this sense, the models are used as the main body of a deep learning architecture that allows understanding the text. Then, it stacks one fully-connected layer to learn to make predictions like detecting entities in the text or converting the text to a vector representation incorporating all the semantic and syntactic information. This vector representation (also called text embeddings) can be used to identify similar texts. To that end, we propose to use BERT to match unstructured information directly with the linked open standards based on the linked data descriptions.

A typical case study for this is the case of recruitment in large EU organisations or public administrations, where recruiters need to search the best candidate for a given position from a huge amount of resume profiles. A typical practice in the industry is to utilise full-text search engines like Elasticsearch, where the documents are indexed to be used for fast keyword search and to limit down the number of retrieved profiles for the recruiters to inspect. However, these engines fail to capture the semantic meaning of the text and thus disregard an important number of relevant profiles resulting in an extensive inspection of the resumes by the recruiters that could also lead to bias due to human fatigue. Another typical practice used by the recruiters is to identify a profile exhibiting similar skills to the job description out of the retrieved job applicants and use an automated approach for identifying similar profiles. Yet, this mainly depends on keyword-based approaches.

To that end, a comprehensible and straightforward way to identify similar profiles is to extract the skills from the experience descriptions, match them to existing linked open data (e.g., ESCO) and use them as input for resume similarity scoring. To achieve this, a typical method is to train a ML model for skills extraction that, however, requires manual labelling which is time and cost consuming.

We propose Resume2Skill-SE (Search Engine) architecture, an unsupervised approach for skills extraction that leverages (a) the BERT architecture and Siamese Networks for mapping the descriptions into a vector representation and (b) external knowledge from the ESCO classification of skills and occupations using the Faiss algorithm of efficient similarity search for scalable and efficient skill search [17]. Furthermore, the architecture uses the matched skills to model a profile-skills bipartite graph that allows calculating similarity score between resumes based on different formulas. This use case indicates that by leveraging linked open data and simultaneously applying ML techniques, we can ensure high data quality through matching to linked open data. This will subsequently lead to semantic interoperability.

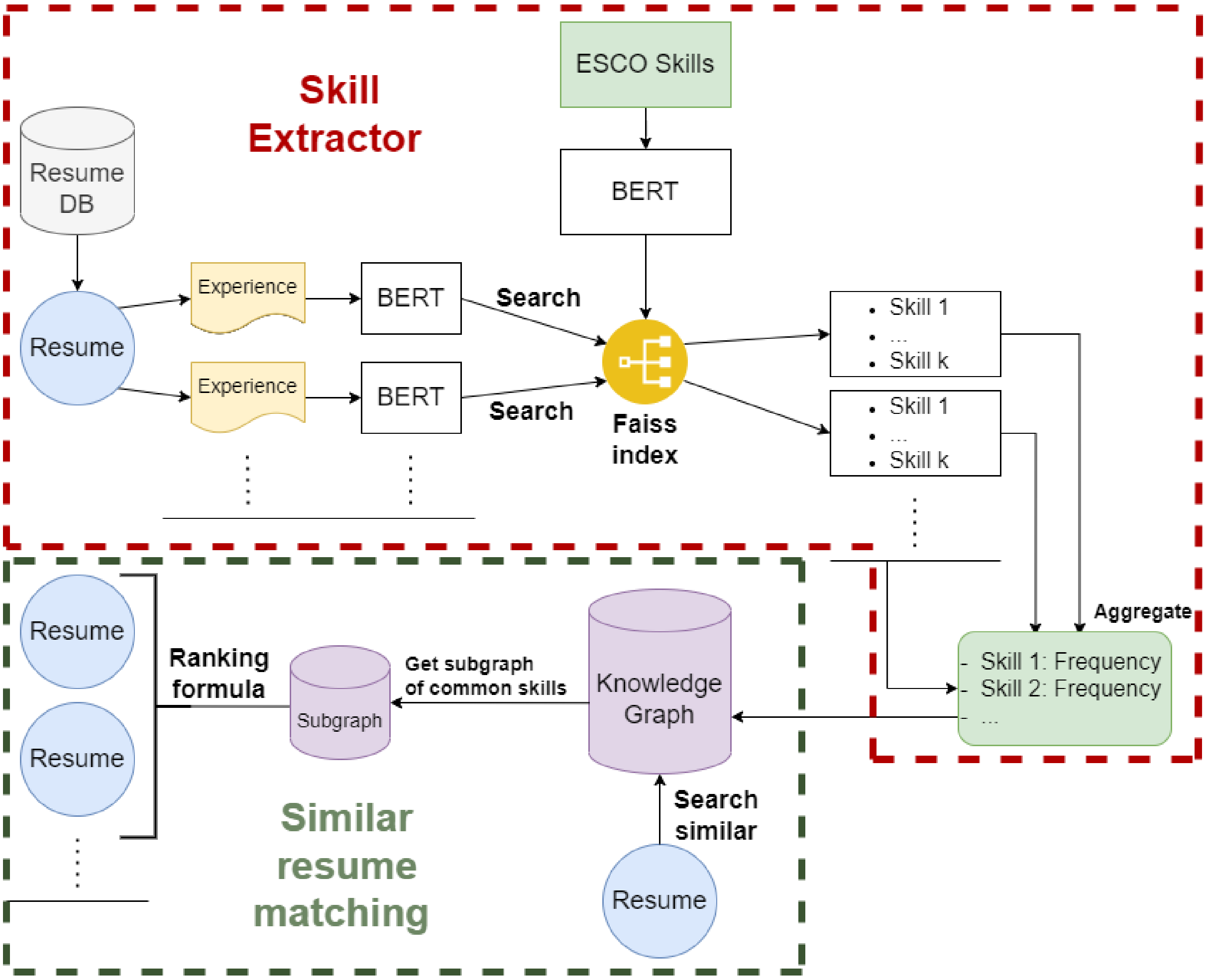


Figure 6: Resume2Skill Architecture

Figure 6 presents the architecture of Resume2Skill-SE. The steps are described as follows:

1. Initiate a graph database with the resume profiles and the ESCO skills without relationships.
2. Create vector representation of the ESCO skills labels and definitions using BERT.
3. Train an index on these vectors by utilising the Faiss algorithm.
4. Collect each resume profile with the list of its experience descriptions.
5. Create vector representation of the experience descriptions using BERT
6. For each experience description, search on the trained Faiss index for the top k relevant ESCO skills.
7. Aggregate the extracted skills per profile by counting the number of experience descriptions an ESCO skill appears. The reasoning behind this is that when a skill appears in multiple experience descriptions then there is a higher confidence that the resume includes the respective skill.
8. Use these aggregations to ingest relationships between profiles and skills in the graph database.

9. In order to identify the most similar profiles from a given profile based on the skills, first get the subgraph of profiles that have at least one common skill with the source profile.
10. Apply the following profile similarity ranking formula:

$$simScore(p_i, p_j) = \bar{P}_{si} \circ \bar{P}_{sj} = \sum_{k,t} P_{si}[s_k] \cdot P_{sj}[s_t], s_k = s_t \qquad (1)$$

where p_i and p_j are a pair of profiles and P_{si}/P_{sj} are the related skills with weights, respectively. In this way, we consider a weighted sum of the target profile weights based on the source profile weights. As a result, if a skill weight of a source profile p_i is high and the target profile has also a high weight for the same skill, it will lead to a high similarity score.

Figure 7 shows an example of the resume profiles-skills bipartite graph.

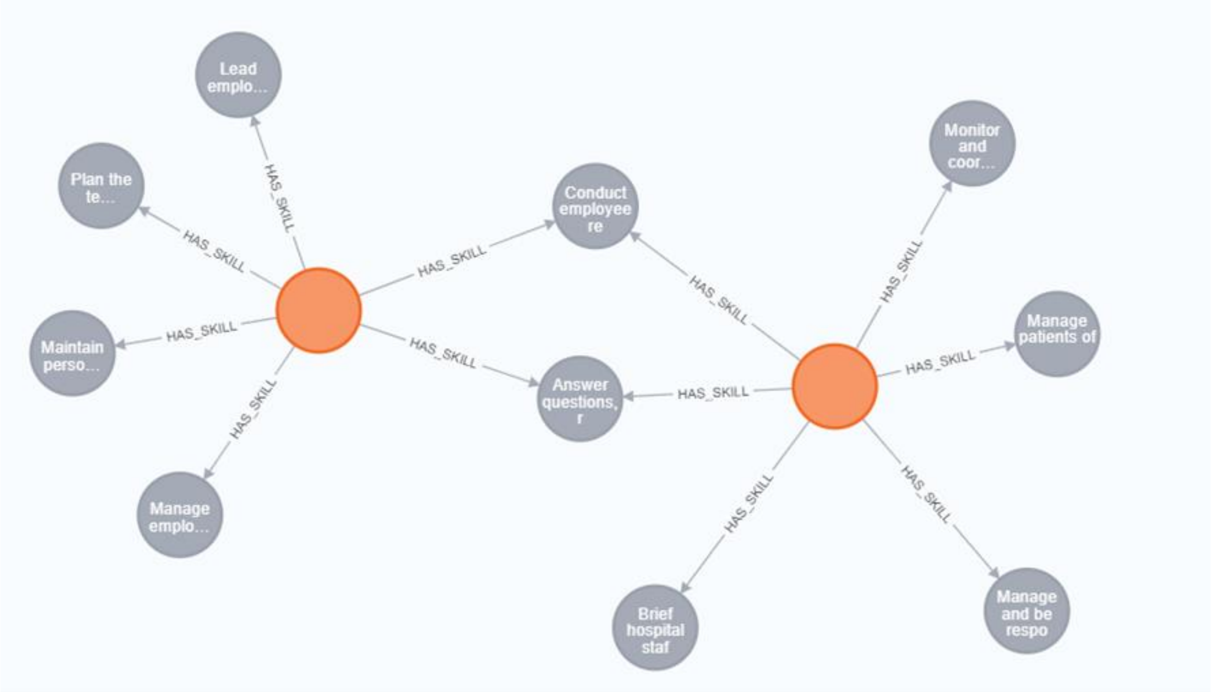


Figure 7: Resume-skills bipartite graph in Neo4j

4.3.3 Recommender system for schemas matching and merging.

When developing new evidence, we have (a) existing ontologies and (b) descriptions of evidences from various countries. The idea of the system would be to merge attributes and relationships (representation as a graph or a tree) between evidences (method would need to support multi-language approach). The merged schema would then need to be automatically matched to existing ontology/vocabulary.
Other use case of using the same techniques would be through a guided approach by manually creating a new evidence type.

4.3.4 Competence Organization Discovery

Evidence issuing discovery helps users to identify the authority with legal competence to issue a specific evidence type, which is particularly required when implementing a cross-border cross-sectoral OOP for public services. The EBSI has ignored the evidence issuing discovery for now; SDG OOTS and DE4A projects are relying upon catalogues to help on the discovery, Data Service Directory and the Issuing Authority Locator respectively. Although the aim and the structure of these catalogues is well known by every project participant, the reality poses a great challenge to maintain such catalogues up to date.

Governments are organised in public units according to a hierarchical structure, where each public unit has a set of responsibilities that constitutes its competences within the government, such as competences on managing base registries and systems to access them. Both government structure with distribution of competences and their changes are lawful when some legal text that establishes such an organisational arrangement is published in the Government Official Gazette. Member States with decentralised governments have several Government Official Gazettes, one per statal, regional and local government, but their gazettes are known and fixed.

As a government organisational arrangement may be amended over a legislature or because a change of legislature, they are likely to change. As any government organisational arrangement that is published in the Government Official Gazette has to be manually transferred to the catalogue that represents the government structure and distribution of competences, such catalogues are hard to maintain and there may be a significant delay between the publication of an arrangement in the Government Official Gazette and the corresponding update of the catalogue. However, organisational interoperability depends on the reliability of these catalogues.

To ensure that catalogues will be updated correctly with the latest amendments, it is clear, that there is a need to avoid manual transfer of information and automate the procedures regarding the processing and consolidation of organizational arrangements. As mentioned before, government organisational arrangements are published in legal texts through the corresponding Government Official Gazette, so these provisions are expressed in natural language. Natural language processing technologies allow systems to understand language spoken and written by humans, thus, NLP can be used both for the analysis of legal texts and the identification and the extraction of useful information in order to convert such provisions into structured data to automatically update the corresponding catalogues.

NLP cannot only help to maintain organisational catalogues updated, but to notify when some relevant competence has been moved to a different public authority thus the usual consequential changes in addresses and systems, such as web domains, can be alerted. Besides, NLP may also help to maintain catalogues updated regarding web portal addresses that can be found in internet associated to each public authority.

Considering that legal texts are typically characterized by a peculiar language, are convoluted and have unnatural syntax, legal knowledge extraction is heavily intertwined with natural language and common sense (i.e., need for general context understanding). In addition to this, there is lack of consistency in terms of the way these documents are structured. This makes accessing the content embedded in legal texts a particularly challenging task, therefore, general-purpose NLP engines may not be the appropriate choice as they ignore the particularities of the field.

To address this challenge there is a need for a thorough analysis of legal texts in order to identify patterns for both the organizational structure and the content of such documents. The identified patterns could help to extract the specific information that we are interested in (e.g., articles that refer to the hierarchical structure, public units, responsibilities) and modelling advanced language rules that will be further utilised by ML techniques and language models.

Diving into this complex area, we figured out a lack of research efforts towards the automated extraction of hierarchical structure and responsibilities of public administration units. Related contributions such as Metadata extensions from the Greek Government Gazette [18] follow rule-based approaches, as they seem to increase system's precision. However, despite the complexity and dynamic nature of legal texts there is a plethora of scientific studies that examine the possibilities to extract quality information using ML methods and NLP techniques.

The most common NLP tasks that are related with the information extraction is the pos-tagging, the dependency parsing, named entity recognition, and entity linking. The first two tasks are cornerstone elements of a NLP pipeline and they are usually part of the rest of the tasks. Named entity recognition is used for identifying specific entities (e.g., public units) in the text, while Entity linking is the linking of these entities with the corresponding entities in a knowledge base. Another important task is the extraction of relations (Relation Extraction) among objects in the text. In order to perform these tasks various ML techniques are used, like neural networks, word embeddings and state-of- the art transformers (e.g., BERT). Example applications of such methodologies on legal texts are:

- The reveal of implicit relations based on approaches for building vector representations of words, such as Word2Vec, FastText, as well as vector representations of texts and sentences: BERT and Doc2Vec. [19]
- Automatic extraction of amendments using neural models (BERT and BiLSTM) [20]
- Conversion of legal text into structured open-data [21][22][23][24][25][26]

Assessing the best practices towards the analysis of legal texts, we understand that extracting and updating the organisational structure and competence of a public unit demands a combination of such methodologies. Therefore, it is expected that a holistic approach could be a proper solution towards this direction.

In summary, publishing up to date structured information on the government organisational structure, competence distribution and web domain addresses helps users to locate public services provided by such a government, including evidence issuing services.

4.4 Lingua franca and automatic language translation

We are in a digitally connected multilingual world, where understanding, mapping and relating to different languages is a mandatory part in integrative and borderless citizen centric e-services. Starting with google translations many solutions for language translations are available up to date. However, automatic translation of documents may pose a challenge for information exchange in public services since automatically translated documents do not have a legal value, and hence cannot be treated as lawfully issued evidence. In a cross-border public service provision process there are two occasions where the language matters for the user, 1) The canonical evidences, 2) The messages when canonical evidence is exchanged (namely the portal pages of the explicit request of the evidence and the page where the evidence is displayed to the user to check and approve). In cross-border information exchange, typical process involves language translation to and from the lingua franca namely (British or American) English. DE4A have solved the multilinguality problem by introducing the Multilingual ontology repository, which is semi-automatic, where the canonical evidences in all the languages in use are translated, stored and maintained in a database, and, an API translates the portal webpages to the desired languages [27]. The problem of legality of translations is tackled by a two-step process of non-verified (machine translated) and verified (verified by a domain expert and approved) labels that is visible in the evidence exchanged. More information on the technical description of MOR can be found in DE4A semantic repository through https://github.com/de4a-wp3/MOR.

Availability of EU official documents in official languages in the EU countries has been a topic in EU initiatives since the early days of digitalisation of public services [28]. Tools have been developed for translating legislations and other legal documents automatically [28]. Tools such as DGT-Translation Memory [29] consisting of a language resource corpus where the domain specific terms in different languages have mapped to each other called “parallel texts”. The mapping at word, expression and sentence level is available for those who want to develop new translated documents. This corpus is playing the role of the translation memory, similar to the “black box” of the Turing’s original experiment of language translation [30]. This approach is also a legally viable semi-automatic translation method. Few other tools consist of parallel texts in different domains, especially matching by pairs of languages, so that sentences of each language are mapped to the other is available within EU translation resources as described in Steinberger et. al. [28].

4.5 User friendly digital systems supported by Chat bots

Alongside of any technology deployment, onboarding the users is one essential activity. Increasing the awareness of the technology, increasing the usability and perceived uses of functions of the system, and other advances, will help reduce the complexity of the technology. Onboarding different customer segments is typically found as a considerable challenge in the change process of e-Government. Different methods exist for customer onboarding, such as focus groups, early adapters, top- down implementation, and so on. However, the time and the cost related to the process may vary based on the customer segments, demand and complexity. The challenge becomes severe when the target group of technology solutions are the general public. Onboarding every citizen using classical training methodologies is time consuming as well as ineffective, due to different levels of digital literacy among the users. Automated solutions for answering the frequent questions become handy in such situations. Chat systems associated with web pages for automatically answering questions or connect to a live Q&A session is not a new technology.

Chatbots come in very handy in helping the citizens, when they come across problems in the process of using the digital systems. Chatbots are simply communication agents - computer programs - that are able to detect and interpret human natural language through text or speech. In doing so chatbots use various text recognition and pattern recognitions tools (algorithms). Early chatbots were just pure text that contain answers to a bunch of frequently asked questions [31]. Current state of the art however progressed way beyond this and use AI, NLP and text mining technologies to answer tailored questions from users, mainly related to functionalities in their services through web and online forms [32]. Although, the authors of [32] report that there is no evidence for change happened to the services nor to the governance system due to introduction of chatbots, the process of information provision may have a had a considerable influence towards the efficiency and reliability of the service delivery. Filling in government forms require knowledge and understanding of the terminology and the meanings of the form entries for accurate information provision. Resolving ambiguity is a key in this sense. A chatbot with advanced technology may be beneficial for such situations [31][32].

The role and the use of chatbots in the context of information exchange across-borders has not yet been explored according to the related literature. However, the current practices have a big potential to be extended to integrative e-services overcoming organisational barriers [31]. A large extent of the challenges may also lie within semantics. A strong natural language parser is typically behind a good chatbot and can: translate voice to text and/or vice versa; maps (understand) the questions to correct answers; speak the same language as the user; and so forth. Also, advanced functionalities are implemented in these tools.

Additionally, these chatbots are progressive tools, where time is needed for model maturity by user interactions. NLP technologies such as name entity recognition could solve ambiguity problems in mapping correct keywords in questions to answers. If a data structure is behind the questions, the definitions to concepts are available, and, efficient search algorithms are driving the search engine, a chatbot could become useful. In cross-border public service, chatbot tools may find additional challenges with language-specific meaning of certain concepts, in addition to the specificity of their local bureaucratic processes. The way forward in that regard can be a similar process we followed for evidence exchange, i.e., semantic agreement of concepts, classes, and attributes between the respective competent authorities in the countries involved.

4.6 Summary and way forward

This chapter pinpointed several ways and possibilities of using advanced technologies, in the light of the state-of-the-art of the digital e-government, and the data that is available from the cross-border public service provision. However, lack of resources and data limited us in proving some of the concepts introduced in this chapter. These concepts could be tested in the future when necessary algorithm inputs are available for testing.

5 Conclusions

This deliverable presents the landscape of using advanced analytics in public administrations, public services, and possibilities to enhance data interchange and interoperability. We provide a comprehensive literature review, possible use case scenarios within the DE4A, and anticipated further developments after the productization of the DE4A.

Chapter 2 draws conclusions based on the systematic literature review (SLR) of the state-of-the-art. A systematic methodology was employed, within which we defined scope, resources, search criteria, and inclusion/exclusion protocols. The main aspects of SLR were "public administration," "artificial intelligence," and the "semantic Web." Out of 399 relevant publications we finished with the selection of 20 most relevant to the DE4A. The exploration analysis shows results of automatic natural language processing analysis to get overall insight into the research area. The more in-depth review showed that the topics are gaining importance (based on number of publications) in the last years. The majority of selected studies are focused towards discussions, applications proposals, different assessment tools or framework definitions for deploying AI for public services. The semantics is mostly presented via specific use cases and as an enabler of interoperability among different institutions and countries. As a result, semantics is required to build efficient IT systems and further upgrade them using AI technologies.

Data is the basis for analytics. For the AI algorithms the input data can be represented as semantic schemas (e.g., automatic alignment of different schemas), content or payload that public services provide (e.g., birth certificates that should be encrypted) or log messages (e.g., systems behaviour). In Chapter 3 we specifically focused on possibilities of log mining approaches. Logs are also a "side product" of Pilots that will be tested within the DE4A. We present different schemas and data that could be automatically processed to detect service failures, possible attacks, anomalies, etc. Production-ready systems already provide comprehensive dashboards to observe results or to be notified in case of an alarm is triggered (i.e., an AI algorithm detected spurious messaging behaviours).

Chapter 4 introduce ideas for possible applications for deployed DE4A platform (i.e., IDK along with the pilots). To enable faster inclusion of additional semantic resources (e.g., driving license certificate) a semantic search engine could search and match existing schema parts defined in general registries (e.g., ISA2 dictionaries) or country-specific lexicons. The search would also need to support cross-lingual domain-specific search that could be achieved using state-of-the art NLP deep neural network models. In contrast, there are many semi-structured documents (e.g., resumes) that could be automatically processed to uncover "hidden" structures and to provide unifying schemas faster. Public services are used by the masses and many questions arise using them. To address the issue, chatbots could help and provide answers from large amounts of textual data. Some use cases of chatbots in public administration were also mentioned in the SLR.

Finally, the semantic and AI technologies are only two of the building blocks that would ease and speed up the evolution of public administration (or other) systems and processes. Before applying these technologies, it is important to have clear vision and regulations in place as this is necessary for sensitive data wrangling which is mostly the case for public services. The OOTS is just one bright example of a regulation that we believe is a step into the right converged direction.

6 References

[1] 'Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement | Annals of Internal Medicine'. https://www.acpjournals.org/doi/full/10.7326/0003-4819-151-4-200908180-00135 (accessed Sep. 19, 2022).

[2] 'Home - EU Vocabularies - Publications Office of the EU'. https://op.europa.eu/en/web/eu-vocabularies/ (accessed Sep. 19, 2022).

[3] 'Code lists - Eurostat'. https://ec.europa.eu/eurostat/data/metadata/code-lists (accessed Sep. 19, 2022).

[4] 'Project Deliverables | Digital Europe For All', *DE4A*. https://www.de4a.eu/project-deliverables (accessed Sep. 19, 2022).

[5] 'Controlled vocabularies - EU Vocabularies - Publications Office of the EU'. https://op.europa.eu/en/web/eu-vocabularies/controlled-vocabularies (accessed Sep. 19, 2022).

[6] 'Glossary | SCOOP4C'. https://scoop4c.eu/glossary (accessed Sep. 19, 2022).

[7] 'Formal ontology - EU Vocabularies - Publications Office of the EU'. https://op.europa.eu/en/web/eu-vocabularies/concept/-/resource?uri=http://publications.europa.eu/resource/authority/asset-classification/c_89b4bdb7 (accessed Sep. 19, 2022).

[8] 'EUR-Lex HTML (EN)'. Accessed: Sep. 19, 2022. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32018R1724

[9] 'Taxonomy - EU Vocabularies - Publications Office of the EU'. https://op.europa.eu/en/web/eu-vocabularies/concept/-/resource?uri=http://publications.europa.eu/resource/authority/dataset-type/TAXONOMY (accessed Sep. 19, 2022).

[10] *Regulation (EU) 2018/1724 of the European Parliament and of the Council of 2 October 2018 establishing a single digital gateway to provide access to information, to procedures and to assistance and problem-solving services and amending Regulation (EU) No 1024/2012 (Text with EEA relevance.)*, vol. 295. 2018. Accessed: Jul. 15, 2021. [Online]. Available: http://data.europa.eu/eli/reg/2018/1724/oj/eng

[11] 'Ontologies - W3C'. https://www.w3.org/standards/semanticweb/ontology (accessed Sep. 19, 2022).

[12] 'About: XML schema'. https://dbpedia.org/page/XML_schema (accessed Sep. 19, 2022).

[13] D. Valle-Cruz, E. Alejandro Ruvalcaba-Gomez, R. Sandoval-Almazan, and J. Ignacio Criado, 'A Review of Artificial Intelligence in Government and its Potential from a Public Policy Perspective', in *Proceedings of the 20th Annual International Conference on Digital Government Research*, New York, NY, USA, Jun. 2019, pp. 91–99. doi: 10.1145/3325112.3325242.

[14] M. Denscombe, *The Good Research Guide: For Small-scale Social Research Projects*. McGraw-Hill Education (UK), 2014.

[15] N. Reimers and I. Gurevych, 'Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks', in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 3982–3992. doi: 10.18653/v1/D19-1410.

[16] C. Papaloukas, I. Chalkidis, K. Athinaios, D. Pantazi, and M. Koubarakis, 'Multi-granular Legal Topic Classification on Greek Legislation', in *Proceedings of the Natural Legal Language Processing Workshop 2021*, Punta Cana, Dominican Republic, Nov. 2021, pp. 63–75. doi: 10.18653/v1/2021.nllp-1.6.

[17] I. Konstantinidis, M. Maragoudakis, I. Magnisalis, C. Berberidis, and V. Peristeras, 'Knowledge-driven Unsupervised Skills Extraction for Graph-based Talent Matching', in *Proceedings of the 12th Hellenic Conference on Artificial Intelligence*, New York, NY, USA, Sep. 2022, pp. 1–7. doi: 10.1145/3549737.3549769.

[18] ‘NER & Metadata Extraction from the Greek Government Gazette’. EELLAK, Google’s SoC 2018, 2018. [Online]. Available: https://github.com/eellak/gsoc2018-GG-extraction/wiki/Description/

[19] D. Devyatkin, A. Sofronova, and V. Yadrintsev, ‘Revealing Implicit Relations in Russian Legal Texts’, in *Artificial Intelligence*, vol. 12412, S. O. Kuznetsov, A. I. Panov, and K. S. Yakovlev, Eds. Cham: Springer International Publishing, 2020, pp. 228–239. doi: 10.1007/978-3-030-59535-7_16.

[20] A. Smywiński-Pohl, M. Piech, Z. Kaleta, and K. Wróbel, ‘Automatic extraction of amendments from polish statutory law’, in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, São Paulo Brazil, Jun. 2021, pp. 225–229. doi: 10.1145/3462757.3466141.

[21] M. Adedjouma, M. Sabetzadeh, and L. C. Briand, ‘Automated detection and resolution of legal cross references: Approach and a study of Luxembourg’s legislation’, in *2014 IEEE 22nd International Requirements Engineering Conference (RE)*, Karlskrona, Sweden, Aug. 2014, pp. 63–72. doi: 10.1109/RE.2014.6912248.

[22] M. Martínez-González, P. de la Fuente, and D.-J. Vicente, ‘Reference Extraction and Resolution for Legal Texts’, in *Pattern Recognition and Machine Intelligence*, vol. 3776, S. K. Pal, S. Bandyopadhyay, and S. Biswas, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 218–221. doi: 10.1007/11590316_29.

[23] O. T. Tran, B. X. Ngo, M. L. Nguyen, and A. Shimazu, ‘Automated reference resolution in legal texts’, *Artif. Intell. Law*, vol. 22, no. 1, pp. 29–60, Mar. 2014, doi: 10.1007/s10506-013-9149-8.

[24] M. Opijnen, N. Verwer, and J. Meijer, ‘Beyond the Experiment: the eXtendable Legal Link eXtractor’, Jun. 2015.

[25] N. Stylianou, D. Vlachava, I. Konstantinidis, N. Bassiliades, and V. Peristeras, ‘Doc2KG: Transforming Document Repositories to Knowledge Graphs’, *Int. J. Semantic Web Inf. Syst.*, vol. 18, no. 1, pp. 1–20, Jan. 2022, doi: 10.4018/IJSWIS.295552.

[26] I. Angelidis, I. Chalkidis, C. Nikolaou, P. Soursos, and M. Koubarakis, ‘Nomothesia : A Linked Data Platform for Greek Legislation’, 2018.

[27] ‘DE4A Multilingual Ontology Repository (MOR) - DE4A’. https://wiki.de4a.eu/index.php/DE4A_Multilingual_Ontology_Repository_(MOR) (accessed Sep. 20, 2022).

[28] R. Steinberger *et al.*, ‘An overview of the European Union’s highly multilingual parallel corpora’, *Lang. Resour. Eval.*, vol. 48, no. 4, pp. 679–707, Dec. 2014, doi: 10.1007/s10579-014-9277-0.

[29] ‘DGT-Translation Memory’. https://joint-research-centre.ec.europa.eu/language-technology-resources/dgt-translation-memory_en (accessed Sep. 20, 2022).

[30] ‘Turing test’. https://dictionary.cambridge.org/dictionary/english/turing-test (accessed Sep. 21, 2022).

[31] P. Henman, ‘Improving public services using artificial intelligence: possibilities, pitfalls, governance’, *ASIA PACIFIC JOURNAL OF PUBLIC ADMINISTRATION*, vol. 42, no. 4. ROUTLEDGE JOURNALS, TAYLOR & FRANCIS LTD, 2-4 PARK SQUARE, MILTON PARK, ABINGDON OX14 4RN, OXON, ENGLAND, pp. 209–221, 2020. doi: 10.1080/23276665.2020.1816188.

[32] C. van Noordt and G. Misuraca, ‘New Wine in Old Bottles: Chatbots in Government Exploring the Transformative Impact of Chatbots in Public Service Delivery’, *ELECTRONIC PARTICIPATION, EPART 2019*, vol. 11686. SPRINGER INTERNATIONAL PUBLISHING AG, GEWERBESTRASSE 11, CHAM, CH-6330, SWITZERLAND, pp. 49–59, 2019. doi: 10.1007/978-3-030-27397-2_5.