# Research and cultural metadata exposed using APIs: steps on the interoperability road

**Nicolaie Constatinescu**
PhD.
Transilvania University of Braşov
Email nicolaie.ciubotaru@unitbv.ro

***Abstract***
*The wealth of data available to the researchers and to the curious citizen seems limitless today. For this reason tools able to retrieve and put together pieces of useful data are in growing need. This study is looking into Application Programming Interfaces as means to establish an ecosystem fostering data interoperability for persons and machines. The research and culture landscape is in a continuous need for accessing to data and metadata. For this reason, a set of the most representative service providers were investigated through the filter of data accessibility, and how the data provided fit the needs of possibly machine-assisted tools. The providers are well established actors in the scholarly communication arena with a host of some from the culture heritage field. A supplementary set of trusted repositories which are recipients of CoreTrustSeal (CTS) were investigated in searching for interoperability extensions via APIs.*

***Keywords****: application programming interfaces, api, research metadata, data sets, interoperability, vocabulary namespaces*

## 1. Introduction

Existing APIs are an integral part of the big systems, sophisticated software implementations that are managing research data and metadata obtained out of the current practice. The data are produced by the memory institutions (libraries, archives, museums), and international scientific information resource providers. According to Crossref in 2022 a number over 7.2 million scientific articles were published (Knoth and Zdrahal 2013) setting a contextual figure that ensues a need for different approaches dealing with research outputs. A blog post (Polischuk 2023) from the same service mentions 140 million metadata records deposited by the end of March. CORE through The Open University, the biggest Open Access scientific research aggregator reports in 2023 that 275 million papers are available to access and use through their services out of which 140 million full text papers. This deluge of scientific research results needs new approaches with regards to access the wealth of information trapped in the works. Researchers need new ways to access the research outputs, and new ways to find valuable and suitable data and information in the growing heap. To this end, Application Programming Interfaces (APIs) are rising as a way to pinpoint and access the right pieces at the right time.

APIs are not the means to an end when data is the target, but means to access useful metadata without using the interface of a particular system.

In order to investigate the existing APIs, a selection of representative data providers were taken under scrutiny coupled with the dataset of the CoreTrustSeal certified data repositories available at https://amt.coretrustseal.org/certificates. The reason behind this coupling in the analysis is driven by the mandate for long term digital preservation, which in turn guaranties also metadata and data preservation with a continued provision for access. The main concept followed in research data European policies is data management life-cycle, and part of this cycle is accessibility. The study looked into what characteristics of the API and what parts of the metadata representation of the digital objects would be useful in terms of easy access, interoperability, and reusability.

## 2. Analysis of the APIs

The aim of the analysis is to determine the general common characteristics, the most used serialization formats and protocols, if the endpoints are documented, what range of identifiers are available belonging to what namespaces, and which are the limitations. The APIs were selected from the scientific research / scholarly communication and cultural heritage domains. The living dataset is available for further investigation at Github: https://github.com/kosson/apis-data-source.

To complete a comprehensive picture, a number of fifty two APIs were analysed seeking to determine what a researcher should expect from the endpoints on the following axis:

- access mechanisms (protocols and standards);
- serialization formats;
- metadata schema used;
- common identifiers used for the entities present in the metadata records;
- access and rate limitations;
- Github/Gitlab presence;
- bulk download availability.

The folowing APIs were taken under investigation: American Archive of Public Broadcasting API, arXiv API Access, Springer Nature API portal, Caselaw Access Project bu Harvard Law School, Congress.gov API, CORE API, Crossref REST API, Dataverse, Digital Public Library of America, Europeana API, HathiTrust Digital Library, IEEE Xplore API, Internet Archive Developer Portal, The Lens, JSON/YAML for LoC.gov, Chronicling America, The New York Times Developer Network, OECD data for developers, Open Researcher and Contributor ID, Joint Research Centre Data Catalogue, data.europa.eu - The official portal for European data, Elsevier Research Products APIs, F1000Research, Clarivate Developer Portal, Wiley Online Library, Crossref Unified Resource API, Data.Bibliotheken.nl, BnF API et jeux de données, Open Data at the BnL, German National Library catalogues, Data Catalog of the National Library of Finland, Museums Victoria Collections, Kungliga biblioteket Library Database API, The OpenAIRE APIs, Semantic Scholar Academic Graph API, { NASA APIs }, OpenCitations, re3data.org, Digital Bodleian Search and Data, OpenAlex API, Europe PMC, DOAJ API, Open Library, WorldCat

Search API, Exlibris Alma REST APIs, Research Organization Registry, Sherpa APIs, DataCite REST API, bioRxiv API, Altmetric, Unpaywall, FAIRsharing API.

These APIs were thoroughly investigated and in cases where documentation did not provided easy to access samples of data, accounts were created under providers' guidance to obtain access tokens for securing downloads.

The following providers went under testing in order to obtain sample data using the well-known Postman[1]: arXiv, Orcid, Springer Nature, FAIRSharing API, DPLA (Digital Public Library of America), New York Times, and Elsevier.

The investigation carried out looked into the possibilities for Text and Data Mining if that possibility was explicitly mentioned by a provider. A particular attention was given to the diversity of the providers seeking a revealing well balanced mix of practices and records variety. In the end, the structures and metadata of the digital object representation would point out what is is possible to do with today existing instruments, and the existing metadata. Open Data produced by the Public Sector Information generating institutions were excluded. This decision was driven by the vast swats of catalogues already existing which would potentially skew the perspective due to the sizes and already well ahead practice homogenization. In this context, several conclusions of the 2020 report: *Application Programming Interfaces in Governments: why, what and how*(Vaccari *et al.* 2020) states the need to introduce APIs as a component of digitization due to their modularity and high degree of reusability. At the European level, the adoption of APIs should be done in a coordinated way to avoid the negative effects that ad hoc adoption brings to long-term exploitation. One aspect well captured in the report is that APIs allow institutions to avoid the data "siloed" effect.

## 2.1. Method used

The data was gathered during five months beginning with February of 2023 ending at the beginning of June. A good balance was sought with regards to the institutions involved in this study. Scholarly communication and cultural heritage fields were chosen as data originating form their APIs has the potential to connect and become pieces of exploitation chains/workflows. All APIs were investigated accessing their institutions' sites, reading the help and FAQs pages. For most of the APIs discovered accounts were created looking into the possibilities of accessing data and metadata directly. After accounts were created for the outlets requiring this approach, a collection was created in specialized software Postman aiming to extract sample data for further analysis. The analysis of these samples was done to find the identifiers used in the data and metadata.
The national and international data portals for public sector information were not included in this study because of the sheer size and the scope of those repositories.

[1] https://www.postman.com/

## 2.2. Access mechanism

The most important aspect sought after by the data savvy researcher is the means to gain access to the data and metadata. As expected, the prevalent means of access of today is through the use of HTTP protocol. And this comes with the following explanation. Into the HTTP request-response cycles we have assimilated all sorts of querying forms. The prevalent model is following a RESTful behaviour.

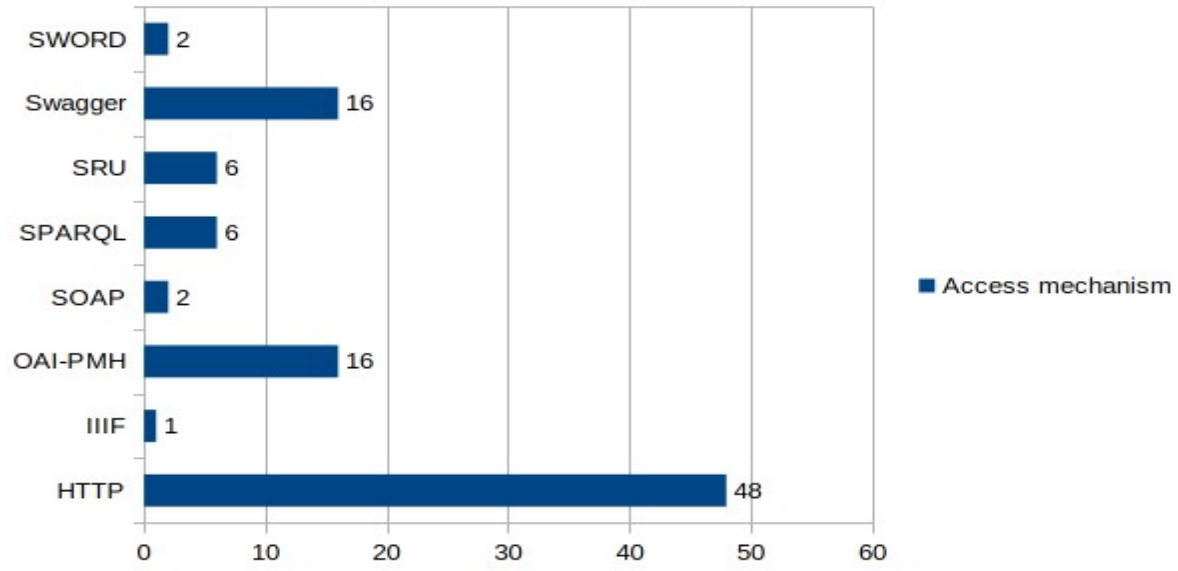


Figure 1: Means of access to data, generically called access mechanism

As seen in the Figure 1, most of the protocols mentioned are using HTTP protocol as subsidiary means of communication, but what is worth mentioning is the diversity of protocols/specifications in use. At times, some of the providers offer multiple means of access to the metadata. Relevant for this variety are mostly cultural heritage institutions like Data Catalog of the National Library of Finland, BnF API et jeux de données or Kungliga biblioteket Library Database API with a notable research outlet: Exlibris Alma REST APIs.

Another aspect pointed out is that the OpenAPI Specification (Swagger) gains traction as an uniform way to document and present the endpoints through which an interested party may access the metadata. The increase in the desire to standardize access is noticeable, as testified by the existing documentation of endpoints using Swagger (https://swagger.io/), that is, adherence to the OpenAPI specification, in fact. This is due to the need to have APIs that behave predictably, benefiting from the best practices of the last twenty years. I chose to introduce also SPARQL (SPARQL Protocol and RDF Query Language)[2], which is also a possible path for querying semantic databases. This option was introduces by the realities of APIs query mechanisms encountered.

[2] https://en.wikipedia.org/wiki/SPARQL

Some of the providers also expose a SPARQL endpoint in addition to RESTful ones. This adds complexity and is prof of existing curated knowledge graphs, a higher form of data aggregation addressing very specialized users. The renown technology analyst Kurt Cagle explained in one of the best articles hosted by Data Science Central(Cagle 2022) what are the powers of the Knowledge Graphs, and how organisations are taking advantage of this technology able to accommodate desired ontologies. There is a potential conclusion here concerning the API relation with the knowledge graphs. Some of the analysed organisation might have their own internal knowledge graphs shaping data according to their own ontologies of their own choosing. The access to external parties is offered through *common APIs representing access to resources that can be styled in specific ways*. In turn this leads to a failed reuse scenario as Mr. Kurt argues that *the interoperability argument falls flat*. The scientist, the curious citizen, and the machines are left to what could be gain out of the metadata exposed.

## 2.3. Serialization formats

A piece of information that proves useful for those searching a means to connect to the wealth of the data available is the format in which they will receive the data. Since the 90s of the last century and then following the widespread adoption of RESTful APIs after 2000, a lot of data serialization formats have been used.

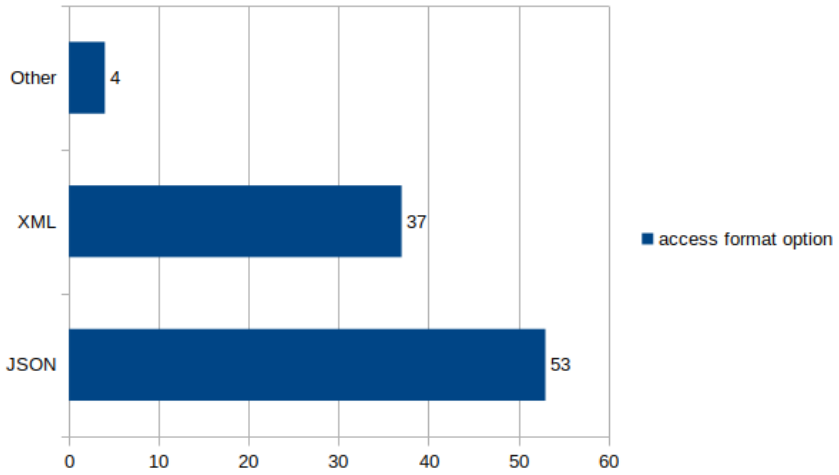


Figure 2: The format distribution as option to access serialized data

JSON format seems to be the format of choice as it is the one present as a choice for serialized data in all the analysed data APIs. XML is the second format present as an option to access the metadata. XML holds value due to the flexibility and already established web standards.

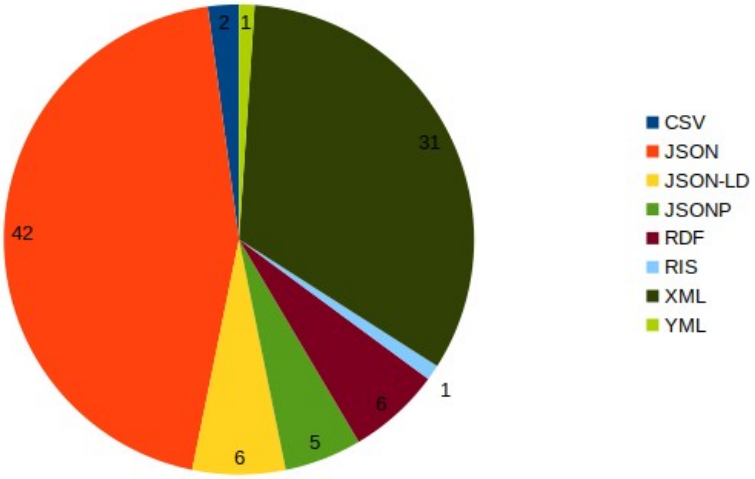


Figure 3: The extended view of the formats available for serialized data.

As evidence suggests (figure 3), JSON formats present some variations in the form of JSON-LD and JSONP in cases where more data needed to form a richer context.

It should be mentioned the fact that for the many of the APIs investigated, the response metadata could be requested in many different formats according to the needs of the caller. This particular aspect is a precious feet due to the versatility and adaptation to the user needs and diverse tooling.

## 2.4. Metadata schema and namespaces used

For the purpose of looking into interaction and interoperability, the samples of data obtained were investigated in searching of existing vocabulary reuse. Out of the lot, only eighteen provided easy identifiable namespaces in their metadata. The rest need a thorough future investigation as the sample data or the extracted ones under the created accounts have not yielded consistent results leading to a hastily conclusion of non-existence.

| **Name** | **Namespaces** |
|---|---|
| arXiv API Access | atom |
| Springer Nature API portal | dc |
| Crossref REST API | rdf, dc, prism, owl, bibo, foaf |
| Digital Public Library of America | dc, dcterms, dcmitype, dpla, ore, rdf, rdfs, skos, owl,edm |
| Europeana API | dc, dcterms, ore, rdf, skos, owl, cc, foaf, rdaGr2, wgs84, edm |
| JSON/YAML for LoC.gov | dc |
| Chronicling America | dc, dcterms, ore, owl |
| Joint Research Centre Data Catalogue | dc, foaf, og, vcard, owl |

| Name | Namespaces |
|---|---|
| Elsevier Research Products APIs | prism, dc |
| Data.Bibliotheken.nl | rdf, owl |
| BnF API et jeux de données | dcterms, rdf, rdfs, skos, foaf, rdaGr2, srw, dc, onix |
| Open Data at the BnL | dc, dcterms |
| German National Library catalogues | dcterms |
| Kungliga biblioteket Library Database API | dc |
| The OpenAIRE APIs | oaf |
| OpenCitations | dcterms, rdfs, owl, foaf |
| Europe PMC | dc, dcterms, dcmitype |
| Open Library | dc |

Table 1: Namespaces used in metadata

The variety of the namespaces involved in the metadata presented in the data is dictated by the profile and the mission of the institution where is considered as valuable to the outside potentially interested parties.

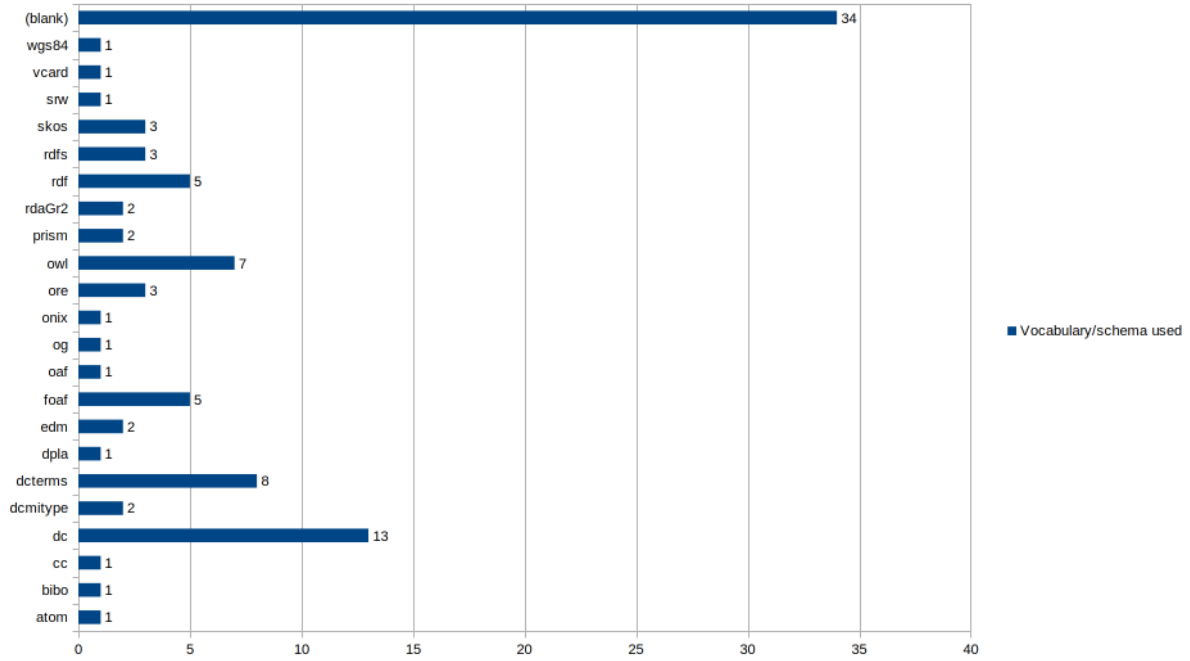


Figure 4: Represented namespaces in the metadata collected

Dublin Core is one of the main provider of vocabulary terms used in metadata construction. We do need to couple the namespaces used also in linked data representation as it is often the case of the providers who also manage a knowledge graph. Beyond the capable interrogation means SPARQL is presenting, when it comes to serialization of data, many of the present namespaces are present (owl, og, rdf).

Although the investigation revealed the namespaces used by most of the APIs, there are some which have not exposed any of the usual namespaces. As Table 1 shows the parties more interested in the reuse of the existing namespaces come from the Cultural Heritage sector. Digital Public Library of America even points out[3] the *Not Invented Here* syndrome that should be acknowledged for the betterment of the data users.

## 2.5. Common identifiers for the entities

One important aspect is addressing which identifiers are present in the records representing the digital objects or the ones that are referenced. An important key in which the following numbers should be read is that all of the records investigated present a localized identifier specific to the system used to produce or manage the metadata. Past this mention, the following figures reflect the identifiers used in the records for the digital objects described.

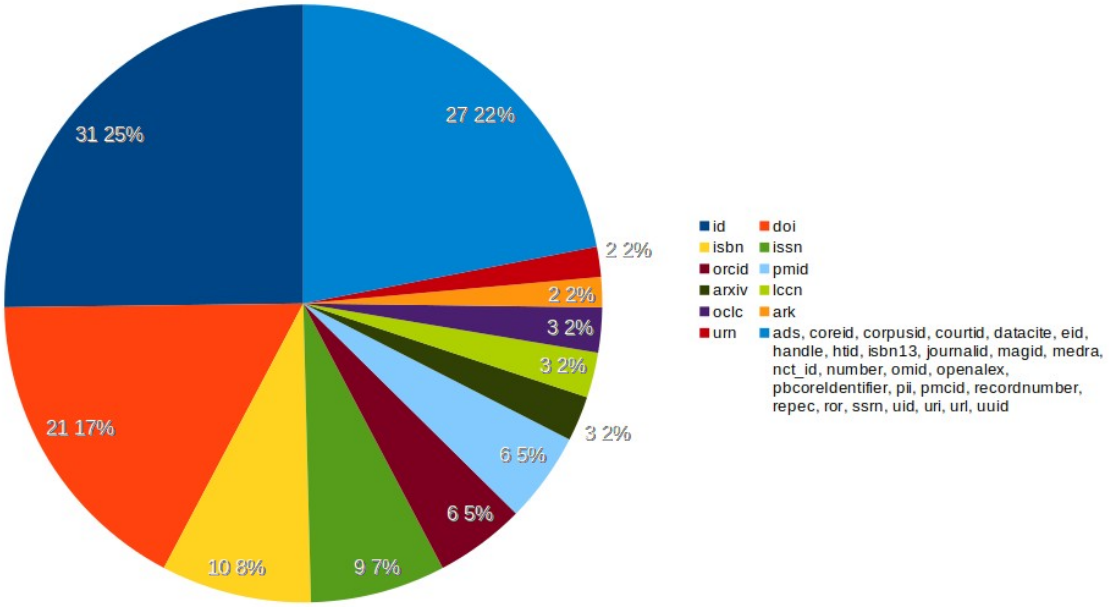


Figure 5: The identifiers present in the record representation of digital objects

Most of the internal ids (*identifiers*) used are alphanumeric sequences generated as a result of a cryptographic computation (hash). This is a perspective of what one is to expect finding in the metadata exposed by an API for a particular digital object described. Many providers are designing their own id name schema, and some have applied cryptography (uuid). Further research is needed to determine if the hashes are computed based on the content or any arbitrary seed content chosen by the creators of the records or is just what the adopted software solution gave by default.

[3] https://pro.dp.la/developers/philosophy

# 3. General details of the APIs in focus

All of the APIs expose a documentation where technical details are combined with useful descriptions of what an interested party is able to find through the endpoints. Some of the institutions go to great lengths including examples of calls on the endpoints involving primary minimal CLI tools like curl.

Some of the investigated API providers are actually building shell software over powerful indexing solutions like Solr or ElasticSearch, even on the base of these sophisticated ones which is actually Lucene. This decisions are factors that model the way requests are built, and this entails prescriptions on how data should be consumed as well.

All APIs give access based on a "recognition token" of sort. Some of them offer their API endpoints access based on an existing account where one should declare the way she/he integrates these with their own application. For the purpose of gathering data, this path couldn't be avoided. In case of The Lens, a 14-day trial access was granted to experience what the service has to offer.
All the API service providers present terms and conditions and most of them set limitations constraining the person or the machine to a certain behaviour in order to preserve the bandwidth and availability.

Although the service providers offer the endpoint, there are some of them who also make periodic data dumps available to download, a convenient strategy to expose all the data at once ready to be parsed and integrated in potential bigger data workflows.

# 4. API data exposure readiness of the trusted digital repositories

The role of the APIs is growing along with the infrastructure used to manage metadata. The infrastructure needs to satisfy the growing concerns on digital long-term preservation. For all the research conducted under Horizon 2020, the researchers need to deposit the papers (*in a machine-readable format*) and data in “trusted repositories” according to the Article 17 of the Model Grant Agreement (Annotated Grant Agreement[4]). This requirement is also permeates into the publication guidelines of Open Research Europe[5], the European Open Access publishing service offered to the researchers involved in Horizon 2020 project.

For this reason, the dataset provided by the CoreTrustSeal has been taken under scrutiny. CoreTrustSeal is a set of requirements for repositories bearing the complete name CoreTrustSeal Trustworthy Digital Repository (TDR) Requirements. This is the collaborative effort of Data Seal of

[4] https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/aga_en.pdf accessed from https://webgate.ec.europa.eu/funding-tenders-opportunities/pages/viewpage.action?pageId=1867974 (Dissemination & exploitation of project results)

[5] https://open-research-europe.ec.europa.eu/for-authors/data-guidelines

Approval (DSA) and World Data System a part of International Science Council (WDS) working together under Research Data Alliance (RDA). This is a tool by which digital repositories are analysed if are abiding to the specifications. As long as all criteria are met, the digital infrastructure is declared able to provide long term preservation services.

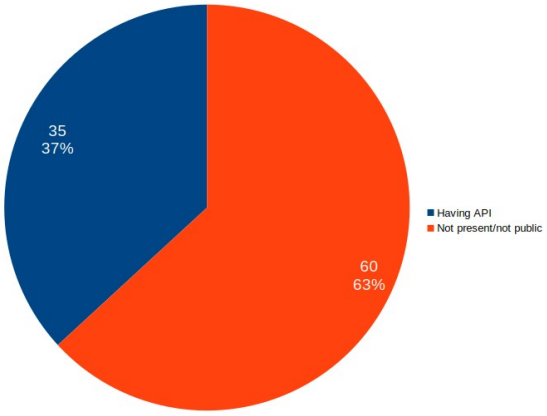


Figure 6: CoreTrustSeal Repositories exposing an API

The data tell an interesting story on the importance of developing and maintaining an API. Almost a third of the CoreTrustSeal repositories have an API. The rest of them have no public endpoints.

Another very important aspect concerning this analysed set is that most of the software solutions are in house developments. This is almost surprising because one would expect more open source or proprietary implementations.

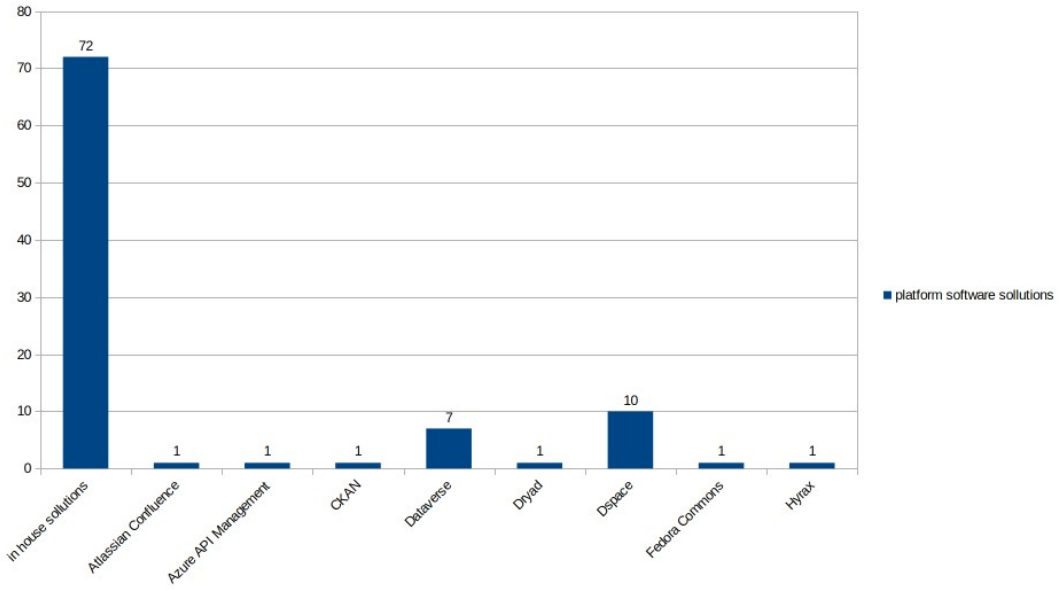


Figure 7: Platform software solutions

The data shows a preference for DSpace and Dataverse integrations for those who went on adopting an open source solution.

# 5. Conclusions

The APIs investigated revealed some common traits and different degrees of existing protocols and policies. All of them are using HTTP protocol to provide access. Twenty years after Roy Fielding proposed model, the APIs are an established means of exposing the data for any interested party, let that be human or machine. The information obtained from the analysis could be the support for those planning to develop a dedicated data access service. The APIs investigated offer access to metadata. The data you may access is actually metadata representation of the actual digital objects for the obvious reasons concerning size and practicalities related to sending large sets over the wire. Even the bulk downloads are consistently metadata data sets. The metadata used to represent the digital objects is varying from outlet to outlet. An integration of these data sources in workflows is hard to obtain due to heterogeneity. Identifiers are minted according to the organisation needs. Digital Object Identifier is the most important identifier adoption wise.

Although Digital Object Identifiers are used on large scale, the permanent identifiers are a secondary means of identification next to the internal id. Most of the internal ids used are alphanumeric sequences generated as a result of an cryptographic computations. This study is also useful for all the digital repositories who follow the REST API first approach.

The Open Science Monitor[6] is an official instrument used to track what changes are occurring in Europe and the world at large. In the section dedicated to *Data on open collaboration*[7] there is an interesting trend to observe with regards to the growth of research APIs.

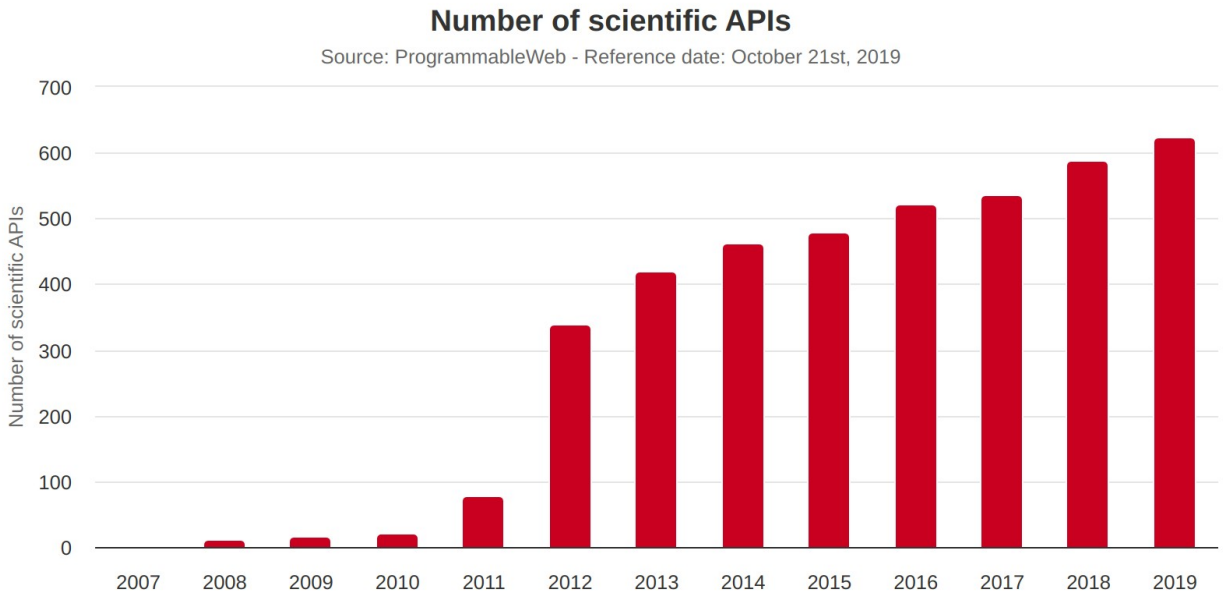

Figure 8: The growth of scientific APIs until 2019 - ProgrammableWeb data

Although, the data was obtained from the now disappeared ProgrammableWeb, a portal where data on existing APIs was kept updated, the growing trend must have been kept at the highest level due to the growing need for data in processes concerning machine-actionable data.

As a further useful development, an investigation into the re3data.org dataset to uncover those who expose APIs[8] according to the criteria developed in this study.

# 6. References

Cagle, K. (2022) From Knowledge Graphs To Knowledge Portals - DataScienceCentral.Com [online], *Data Science Central*, available: https://www.datasciencecentral.com/from-knowledge-graphs-to-knowledge-portals/.

Knoth, P. and Zdrahal, Z. (2013) 'CORE: aggregation use cases for open access', in *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '13, New York, NY, USA: Association for Computing Machinery, 441–442, available: https://doi.org/10.1145/2467696.2467787.

6 https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science/open-science-monitor_en

7 https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science/open-science-monitor/data-open-collaboration_en

8 https://www.re3data.org/metrics/apis

Polischuk, P. (2023) 2023 Public Data File Now Available with New and Improved Retrieval Options [online], *Crossref*, available: https://www.crossref.org/blog/2023-public-data-file-now-available-with-new-and-improved-retrieval-options/.

Vaccari, L., Posada, S.M., Boyd, M., Gattwinkel, D., Mavridis, D., Smith, R., Santoro, M., Nativi, S., Medjaoui, M., Reusa, I., Switzer, S., and Friis-Christensen, A. (2020) Application Programming Interfaces in Governments: Why, What and How [online], *JRC Publications Repository*, available: https://doi.org/10.2760/58129.