Diachronic Linked Data: Towards Long-Term Preservation of Structured Interrelated Information

Sören Auer
Universität Leipzig
auer@informatik.uni-leipzig.de

François Bancilhon
Data Publica
francois.bancilhon@data-publica.com

Peter Buneman
Univ. of Edinburgh
opb@inf.ed.ac.uk

Vassilis Christophides
Univ. of Crete & FORTH-ICS
christop@csd.uoc.gr

Theodore Dalamagas
IMIS / RC "Athena"
dalamag@imis.athena-innovation.gr

Giorgos Flouris
FORTH-ICS
fgeo@ics.forth.gr

Dimitris Kotzinos
TEI of Serres & FORTH-ICS
kotzino@ics.forth.gr

George Papastefanatos
IMIS / RC "Athena"
gpapas@imis.athena-innovation.gr

Helen Parkinson
EMBL - EBI
parkinson@ebi.ac.uk

Dimitris Sacharidis
IMIS / RC "Athena"
dsachar@imis.athena-innovation.gr

Yannis Stavrakas
IMIS / RC "Athena"
yannis@imis.athena-innovation.gr

Kostas Thiveos
Intrasoft International
Kostas.Thiveos@intrasoft-intl.com

ABSTRACT

The Linked Data Paradigm is a promising technology for publishing, sharing, and connecting data on the Web, which provides new perspectives for data integration and interoperability. However, the proliferation of distributed, inter-connected linked data sources on the Web poses significant new challenges for consistently managing the vast number of potentially large datasets and their interdependencies. In this article we focus on the key problem of preserving evolving structured interlinked data. We argue that a number of issues, which hinder applications and users, are related to the temporal aspect that is intrinsic in Linked Data. We present three use cases to motivate our approach, we discuss problems that occur, and propose a direction for a solution.

Keywords

Linked Data lifecycle, Data preservation, Data provenance, Data evolution.

1. INTRODUCTION

There is a vast and rapidly increasing quantity of scientific, corporate, government and crowd-sourced data published openly on the emerging *Data Web*. Open Data[1] are expected to play a catalytic role in the way structured information is exploited in the large scale, and offer a great potential for building innovative products and services that *create new value from already collected data.* Open Data are also expected to foster *active citizenship* (e.g., by means of data journalism, smart mobility, and in monitoring greenhouse gas emissions, food supply-chains) and *world-wide research* according to the "fourth paradigm of science" [12]. The most noteworthy advantage of the Data Web is that it records facts rather than documents. These facts become the basis for the discovery of new knowledge, which is not derivable from any individual data source, and thus help solving information needs that were not originally anticipated by their creators. In particular, *Linked Open Data* (LOD), a term referring to open data published according to the *Linked Data Paradigm* [2] are essentially transforming the Web from a document publishing-only environment into a *vibrant information ecosystem* where yesterday's passive readers have become active *data aggregators* and *generators* themselves.

Given that data-aware practices have a huge potential to create additional value across several sectors [14], it is quite surprising that so little attention has been devoted to *long-term accessibility* and *usability* of *high volumes of data*. Recent studies on the evolution of the Semantic Web [7] and Linked Data [6] reveal that the LOD Cloud[2], a monitored fragment of the Data Web, is subject to frequent changes under no centralised administration. Rarely do datasets completely disappear. More often they *evolve without any indication*, subject to changes in the encoded facts, in their structure or the data collection process itself. In this respect, several challenges arise when preserving Linked Open Data:

- How can we *monitor* changes of third-party LOD datasets released in the past (the *evolution tracking* problem), and how can ongoing data analysis processes consider newly released versions (the *change synchronization* problem)?
- How can we *understand the evolution* of LOD datasets w.r.t. the real world entities they describe (the *provenance*

[1] Several examples can be found at data.worldbank.org, data.un.org, thedatahub.org, datacatalogs.org, open.mflask.com

[2] lod-cloud.net

WOD '12, May 25 2012, Nantes, France.

problem), and how can we repair various data imperfections, e.g., granularity inconsistencies (the *curation* problem)?

- How can we *assess* the quality of harvested LOD datasets in order to decide which and how many versions of them deserve to be further preserved (the *appraisal* problem)?
- How can we *cite* a particular revision of a LOD dataset (the *citation* problem), and how will we be able to retrieve them when looking up a reference in the form in which we saw it – not the most recently available version (the *archiving* problem)?
- How can we *distribute preservation costs* to ensure long-term access even when the initial motivation for publishing has changed (the *sustainability* problem)?

Applying the standard digital preservation practice [1] to LOD, we would obtain individual datasets that are "pickled" and "locked away" for future use. Instead, we advocate a different paradigm. *LOD should be preserved by keeping them constantly accessible and integrated into a larger framework of open evolving data on the Web*. This approach calls for effective and efficient techniques to manage the *full lifecycle of LOD*. In essence, it requires enriching LOD *with temporal and provenance annotations*, which are produced while tracking LOD re-use in complex value making chains [5]. According to this vision both the data and metadata become *diachronic*, and the need for third-party preservation (e.g., by memory institutions) is greatly reduced. We expect that this paradigm will contribute towards a really *self-preserving* Data Web or Data Intranets.

It is worth noting that the bulk of LOD research efforts is focused on scalable RDF data stores and efficient SPARQL query engines in centralized and distributed settings (see [9] for a recent survey) as well as on automated methods for ontology matching and alignment [10]. Managing the *full lifecycle of evolving LOD* requires delving into a wide range of techniques, ranging from data extraction, transformation and integration, to change monitoring, quality assessment and repair, up to synchronization and long-term preservation. The study of these technologies is expected to foster sustainable LOD ecosystems by improving decision making, ensuring transparency in data processing, adopting common policies to privacy-aware data sharing, curation and preservation policies, while minimizing rework.

The remainder of this article is organised as follows: In Section 2, we provide an overview of the characteristics of LOD published on the Data Web or on Data Intranets. In Section 3, we describe three motivating use cases that would benefit from diachronic LOD. In Section 4, we present the main features of the proposed framework, and we review related work in Section 5. Finally, we conclude the article in Section 6.

2. THE DATA WEB AND LOD

If the world's knowledge is to be found on the Web, then we should be able to use it to answer questions, retrieve facts, solve problems, and explore possibilities. This is qualitatively different than searching for documents and reading them, even though text search engines are getting better at helping people perform such tasks. Many major scientific discoveries and breakthroughs have involved recognizing the connections across domains or integrating insights from several sources. These are not associations of words; they are deep insights that involve the actual subject matter of these domains. This is the objective of the *Data Web* (see Figure 1) which extends the current Web infrastructure with a *global data space* connecting data from diverse domains: "a Web of things in the world, described by data on the Web" [2]. We are currently witnessing a smooth transition on the Data Web where published data progressively become more and more powerful, easier for people to understand and use. According to the W3C quality star scheme[3] we can distinguish data:

★ Available on the web (in whatever format) but with an open license, to be Open Data
★★Available as machine-readable structured data (e.g., excel vs. image scan of a table)
★★★ as (2) plus non-proprietary format (e.g., CSV instead of excel)
★★★★ as (3) plus using open standards from W3C (RDF[4] and SPARQL[5]) to identify things through de-referenceable HTTP URIs, to ensure effective access
★★★★★ as (4) plus establishing links between data of different sources.

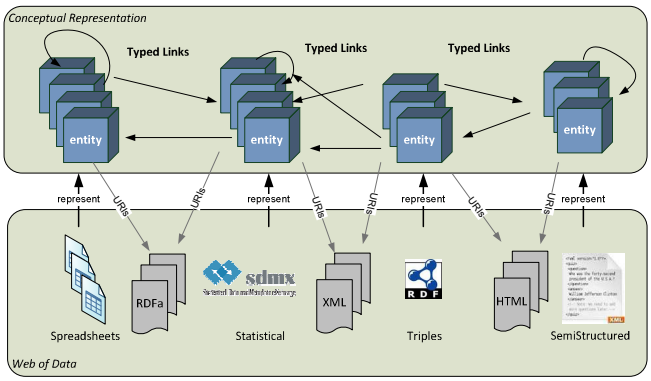

Figure 1: The Data Web

Linked Open Data [4] refers to the recent W3C efforts for a unifying, machine-readable data representation infrastructure that makes it possible to semantically access and interlink heterogeneous resources at data level — independently of the *structure* and the *semantics* of the data, who created them, or where they comes from. The core idea of LOD is to use HTTP URIs to not only identify Web documents, but also to identify arbitrary real-world entities or things. Whenever a Web client resolves one of these URIs, the corresponding Web server provides the description of the identified entity by means of a collection of RDF triples. These datasets may contain links to entities described by other data sources. Links take the form of RDF triples, in which the triple's subject is a URI in the namespace of one server, and the triple's object is a URI in the namespace of the other. The triple's predicate URI determines the *type* of the link. Whenever an application resolves a predicate URI, the corresponding server responds with a RDF Schema (RDFS[6]) or Web Ontology Language (OWL[7]) definition of the link type. These descriptions can in turn contain links pointing at other vocabularies, thereby defining mappings between related

[3] www.w3.org/DesignIssues/LinkedData.html
[4] www.w3.org/RDF
[5] www.w3.org/TR/rdf-sparql-query
[6] www.w3.org/TR/rdf-schema
[7] www.w3.org/TR/owl2-overview

vocabularies. Exhibiting a higher degree of interoperability than documents and ease of reuse, Linked Open Data emerges as a prominent choice for sharing (semi-) structured data worldwide.

In this work, we advocate a *6-star LOD quality*, called *Diachronic LOD*, which enhances data *with temporal and provenance annotations* capturing LOD production and (re-)use context. Diachronic LOD will significantly increase the *functionality* and *meaning* of data published on the Web. This satisfies the need for numerous value-added decision support and business intelligence applications that operate on top of an unbound, global data space and rely on extensive data repurposing and collective intelligence. In this context, understanding how some piece of data was created or where it was copied from, is crucial to assess the *data quality* and strengthen *data accountability* [15]. As a matter of fact, the LOD value chain [5] involves a series of data *stewards*[8], *custodians*[9], *and curators*[10] actually producing, consuming and brokering data products, and is far more complex than in traditional enterprise or scientific contexts. A recent snapshot of the LOD Cloud currently comprises more than 50 Billion facts[11] while as reported in [6] half of the so-represented entities had a change frequency of less than a week. Therefore, it becomes apparent that existing open preservation frameworks proposed for scientific and cultural data [1] cannot cope with the intrinsic features of LOD, which introduce a number of new challenges:

- *LOD are Structured*: Unlike documents, we need to manage not individual facts but entire LOD datasets representing real-world entities for which additional constraints (e.g., name uniqueness) may hold. Moreover, LOD may be interconnected through typed links when they refer to the same or related real-world entities. This calls for effective entity recognition and co-reference methods to rank LOD datasets according to their quality for guiding crawling and appraisal. It also stresses the need for preserving an entire network of interconnected LOD datasets that may prove to be useful for future analyses.
- *LOD are Dynamic*: Unlike closed settings in which data changes are communicated via notification mechanisms, LOD evolution in the Data Web can only be intermittently observed through crawling. In addition, high-level tools are required to understand the changes of evolving LOD datasets and repairing their potential inconsistencies as new real world entities are described or old ones are proven to be erroneous or even become obsolete. In particular, discovering LOD differences (deltas) and representing them as first class citizens with structural, semantic, temporal and provenance information is vital in various tasks such as the synchronization of autonomously developed LOD versions, or visualizing the evolution history of a particular LOD dataset.
- *LOD are Uncertain:* As LOD usage is generalized, their quality may be compromised by various *data imperfections* (e.g., impreciseness, unreliability incompleteness) due to fundamental limitations of the underlying data acquisition infrastructures, the inherent ambiguity in the domain of interest, or even when privacy-preserving applications modify data by adding perturbations to it. Similarly, when LOD are produced by extracting structured information from text, or by entity resolution algorithms in sensor and social data, the results are approximate and uncertain. *Uncertainty* is a state of limited knowledge, where we cannot discern which among alternative statements are true. In this respect, representing declaratively uncertainly and answering queries over probabilistic RDF graphs is a challenging problem not yet related to long-term LOD interpretability.
- *LOD are Distributed*: By definition LOD production, processing and consumption are activities distributed among several actors worldwide. Today, by archiving at remote sites, we have reasonable methods for protecting our data from physical destruction, but this is no guarantee against the economic collapse of the organization that maintains the data. There has been a proliferation of data centres over the past years — many dedicated to the storage of research data gathered at public expense — but one wonders whether, by analogy with early libraries in human history, we are endangering our data by placing it in such centres without replication. The remote sites depend on continued public funding and there are signs that such centres are no more sustainable than early libraries. For this reason, we plan to investigate distributed replication of LOD enhanced with diachronic (temporal and provenance) annotations, and thus make the Data Web really *self-preserving*.

In our view, we need open specifications and tools for preserving and providing diachronic linked data that involve actors from the entire value chain of linked data. The preservation policies defined by producers, and the data needs specified by consumers should be taken into account by third party agents providing linked data preservation services. This cycle of data production – matchmaking – consumption – preservation will maximize the use of the information and the benefits coming out of it.

3. MOTIVATING USE CASES

In this section, we focus on three representative application scenarios, namely *open data marketplaces*, *enterprise data intranets* and *scientific linked data*. These scenarios that feature complementary, yet challenging requirements for managing the lifecycle of LOD w.r.t. (1) the nature of the data (i.e., factual, categorical), (2) their inherent structure and semantics (i.e., from flat relational data to full-fledged RDF graphs), as well as (3) the complexity of the change languages required to understand data evolution.

Open Data Marketplaces: The mass of data created by governments, international bodies, and NGOs represents a wealth of information for the data-driven economy that emerges. During the last years an increasing number of institutional[12] and administration sites[13] in the USA and Europe share Open Data on the Web: only Data.gouv.fr lists 350,000 datasets, while Data.gov.uk currently has 8,200 datasets. In a recent study, Data

[8] en.wikipedia.org/wiki/Data_stewards
[9] en.wikipedia.org/wiki/Data_custodian
[10] en.wikipedia.org/wiki/Data_curation
[11] www4.wiwiss.fu-berlin.de/lodcloud/state

[12] RDFabout.com/demo/census, ckan.net/dataset/cia-world-factbook, ontologycentral.com/2009/01/eurostat/
[13] data.gov, data.gov.uk, data.gouv.fr, data.gov.gr or recently data.eu

Publica and INRIA surveyed the PSI data produced in France and counted 6.5 million files available. Out of these, 175,000 were actual structured data under the form of tables. In parallel, Internet Memory[14] performed in 2010 a complete crawl of PSI in the UK and identified more than 400,000 tables. Such Public Sector Information (PSI) can be exploited in several ways (see Figure 2). Our focus is on the preservation of Open Data in Marketplaces that can be exploited:

- *As a source to produce new datasets*. In this case, data sources are identified and then selected. Subsequently, data are extracted from the data sources, transformed (mined, translated, restructured, enriched, de-duplicated, classified, etc.), and finally delivered on a custom basis. Data are produced as a live object and delivered on a subscription basis. Most customers want a movie and not a static picture. The production of these data is either provided by the actual user of the datasets (this is the case of companies such as

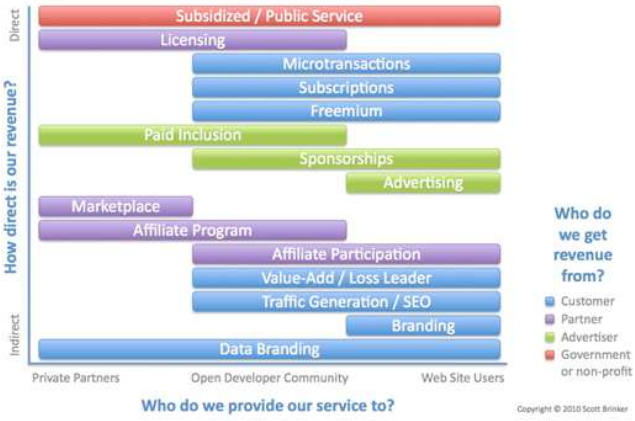

Figure 2: Business Models for Linked Data Publishers

  Reuters[15], Guardian[16], Altarès[17]), or it can be provided as a service by a data editor (this is the case of Data Publica)

- *In Data Portals*, where one can find either the data itself or a reference to the dataset. Various quality enhancement tools are provided: classification, tags, meta data, indexing and full content search, visualization etc. These portals can be operated by government organizations, local bodies (such as municipalities), private companies providing value-added services (such as Data Publica or Datamarket), or citizens groups (such as Open Knowledge Foundation[18] or Regards Citoyens[19]). Finally, one can find directories of data portals that aggregate the content of others.

In both settings, managing the lifecycle of data is important for the analysis of data: e.g., if we want to analyse the correlation between two parameters, we need to make sure we are comparing similar things. In the former case, the main challenge is to take into account PSI data evolution in the new dataset they deliver. Because the data is delivered in a continuous and regular manner, it is necessary to understand and monitor changes so as to take them into account in the final result of the production process. In the latter case, it is essential to have a good grasp and understanding of the public data portals index. Either when they only deliver an instantaneous version of the currently available public data, in which case they need to understand how to monitor changes and retrofit them on their current directory, or when they actually integrate the history and evolution of the public data they index or store, in which case they need to interpret changes and take them into account. Besides establishing a temporal context, public data may additionally be spatially related. To ensure reliability of the performed data analyses, a crucial step is to understand the meaning of change. We need to be in the position to distinguish between changes in data due to changes of facts, data models, and in the data collection process itself. Changes in the data can be of various nature.

- *Changes in the facts themselves*
  - Existing data has been complemented by new data (e.g., we have a new element in a time series).
  - Collecting the budget of an organization over time and making comparisons. In this case, we need to collect two different, seemingly independent pieces of data and put them together in a common series.
  - We store addresses of organizations, and when they move, we can build data of the move of those organizations.
- *Changes in the data collection process*
  - Existing data has been corrected (initial evaluation of a parameter has been revised, which typically happens when the initial estimation of the growth of the economy is revised). This change in data is actually data in itself that should be stored as pertinent information.
  - New data is added (e.g., we are now collecting more data about a fact, or we are listing more or less categories of unemployment).
  - The structure of the data is changing (e.g., we are now collecting data on new geographical zones, or we used to collect data about regions and we are now doing it about departments).
  - The rule for computing data has changed (e.g., inflation rate or unemployment rate is computed according to a different algorithm), and we need to be able to retrofit the new rule on previous data or the old rule on future data.
  - License associated to the data has changed (e.g., it used to require payment and is now free).
  - Correction of the data, in the case of crowdsourcing: we produce an initial data set, and then a crowdsourcing process improves on it, and the data evolves over time.
- *Changes in meta data*
  - Need to find the right meta-data together with the data to understand the current value of the data
  - Need to store the "first derivative" information: when data is replaced by new data, explain the meaning

By enhancing long-term usability of harvested data, we increase the attractiveness of the emerging data markets and we increase the efficiency of the data analytics applications build on top.

[14] internetmemory.org
[15] customers.reuters.com/Home/RMDS.aspx
[16] www.guardian.co.uk/data
[17] www.altares.fr
[18] okfn.org
[19] www.regardscitoyens.org

Table 1: Overview of requirements for the different use case scenarios.

Requirements	**Open Data Markets**	**Enterprise Data Intranets**	**Scientific Information Systems**
Ranking datasets (or parts thereof)	For expanding marketplaces by re-using data from existing pools.	For integration with the enterprise knowledge bases.	For prioritization of external biomedical datasets to be integrated in the curation platform.
Crawling datasets	Intensive crawling of the Data Web for public datasets according to temporal or spatial completeness criteria. Systematic gathering of data about a fact, so as to collect information about the fact in questions.	Acquisition of critical for the analysis structured data and knowledge in a variety of data formats, including HTML tables, files attached to pages and LOD datasets	Identification and acquisition of subsets of public data open data for use in the curation process and meta analyses, and for biomedical data integration.
Diachronic citations	Being able to gather data about a single fact coming from several sources, so as to enrich and improve the quality of the data.	In large enterprises or value chains, datasets are often modified by a number of different departments or stakeholders. Being able to refer to a particular revision is crucial for maintaining data integrity.	For biomedical data, multiple published analyses may exist, which add to, or contradict with existing information. Linking sample annotations to both the source data, and subsequent updates, literature, will allow us to identify conflicts, clarify data production and usage context.
Temporal and provenance annotations	Systematic time stamping of harvested data in order to disambiguate between the situation where "the world changes" and that where "the data about the world changes".	Identify parts of the integrated external datasets, which have changed between subsequent preserved versions.	Biomedical data is temporally structured and refer to adjacent samples, while tracking original vs. curated data versions is critical for their use.
Cleaning and repairing	Three kinds of changes usually occur: those that improve the quality (finer granularity, more information, more attributes),those that degrade the quality (for instance when the census moves from a systematic collection of data to a polling method) and those which entails a different way of computing an indicator. Thus, it is essential to be able to align of the new on the previous structures.	It is a crucial need to maintain the coherence of the enterprise knowledge bases when migrating links and fused information to updated data.	Biomedical data and ontologies used in annotations may become inconsistent w.r.t. to both external datasets and ontologies, as terms become obsolete, change definition, are split or merged. Such annotations must be cleaned and repaired when the external sources they depend on evolve.
Change recognition and propagation	For the data that changes over time, estimate the rate of change (by just testing the data), then ping scan it at a frequency close to that of change.	Re-applying changes, which have been made to an earlier version of a dataset to a newer one and provision of the corresponding conflict resolution strategies.	An automatic detection of changes in external datasets and ontologies consumed internally during curation and meta-analysis would alleviate the curation workload and speed data production/release cycles.
Multi-version Archiving	Depending on the semantics of the data, either datasets are updated by keeping the last collected value, or the sequence of data observed over time is stored. In analogy with accounting practices, we never physically erase anything, but just add a correcting transaction.	Systematically preserving the state of enterprise knowledge bases at certain points in time is crucial for compliance with data auditing regulation policies.	Access to successive LOD versions of reference knowledge made available by scientific authorities, and backward/forward navigation in time between these versions to enable re-analysis and scientific publications.
Longitudinal querying	Respond to queries not just about past states of the data but also time-traveling queries that span across multiple snapshots of datasets.	Being able to query the data evolution is crucial for deriving predictions and forecasts.	To examine data evolution during curation, detect errors and improve curation processes.

Enterprise Data Intranets: We are currently at the verge of an era, where large enterprises are not only adopting the Linked Data paradigm for the integration of their thousands of distributed information systems, but they will also aim to establish reference Enterprise Knowledge Bases (similar to what Freebase[20] by Google) as *hubs and crystallization points* for the vast amounts of enterprise structured data and knowledge, which enables the establishment of a *Data Intranet* completing existing *Document-oriented Intranets*.

In real scale Enterprise Data Intranets the major challenges are related to (1) the *discovery and crawl* of various data types, whatever their format, together with provenance and context information, (2) the *assessment of the quality* of the linked data harvested in the corporate intranet or the Web w.r.t. the data analysis needs of the enterprise, (3) the continuous *improvement of the quality* of the linked data integrated (usually through copying) with the enterprise information systems, and (4) *the monitoring of the evolution* of the open linked data and the *synchronization* of the detected chances with their copies integrated with the private enterprise data.

Scientific Linked Data: Although LOD have been developed in almost all branches of science and scholarship, their use is probably most widespread and advanced in the life sciences[21], particularly, to facilitate community annotation and interlinking of both scientific and scholarly data of interest. Traditionally scientific data in the life sciences has been deposited in international repositories or technology specific databases. Nowadays, they are published on the Data Web and represent a critical resource used by academia and industry in the development of drugs, translational medicine, etc. A great part of the LOD Cloud in Life Sciences is provided by the Bio2RDF[22] platform containing data from around 50 sources. We should also mention that the IMI project OpenPhacts[23] has already identified a collection of open Life Science datasets worth more than 40 billion RDF triples plus 75 business questioned formulated by clinical an pharmacological organizations (such as Pfizer, AstraZeneca, GlaxoSmithKline).

As Life Sciences LOD are consumed, integrated and analysed, the data sources underneath them change constantly though data curation. Curation may involve "manually" reading journal articles or browsing remote databases to find relevant new information. Data gleaned from journal abstracts or copied from other databases and ontologies is typically entered directly by the curator using a Web form or custom interface. Curation also includes automated error checking, and correction, via scripts or fully-fledged applications. This process is what distinguishes curation from related activities such as those performed in data warehouses. Curated information is generally of higher quality, but is correspondingly more expensive to produce and more important to publish and preserve.

Managing the lifecycle of such curated data is critical, and exploiting them can provide accurate results to important biomedical questions. We should also note that data evolution in Life Sciences is even more complex if we also take into account LOD dependencies from various ontologies, scientific databases and dedicated dataflows having in turn their own lifecycles. For example, analysis of RNA-seq gene expression data relies on annotation of the samples used, the software used to align the resulting sequences to the genome, the version of the genome used for alignment and the method and count software used for quantitation. Genome versions are subject to minor annotation changes, or major re-assembly and annotation changes, which means that the entire workflow of analysis needs to be re-run in order to keep the version current. In addition, ontologies employed to map and also structure data may evolve. Changes of external ontologies have to be propagated locally, which may lead to potential LOD inconsistencies. The state of the art in provenance for bioinformatics data has typically been to state the version and/or date of any resource used in publications. As databases were traditionally used to store and manage these data, releases were periodic and publication models were essentially static. As large-scale analyses are performed using the Data Web, this model is now insufficient, and the process of curation and the modelling of data evolution are essential to comprehend and reuse the data expressed on the Web.

In Table 1, we provide an overview of requirements for each application scenario.

4. A FRAMEWORK FOR DIACHRONIC LINKED DATA

To address the needs of the previously described application scenarios, we propose a *distributed, service-based* infrastructure for *curation* and *preservation* of LOD through their entire lifecycle. Such a system will need to comprise the following essential functionality.

Adaptive focused crawling. Gather linked data from the Web about a domain, together with relevant background information that is required to put the data in context. The crawler will take into account the "preservation policies" provided by the data producers, will make decisions on which links to follow first, and will dynamically adapt its frontier accordingly.

Change detection. Identify changes by pulling out and comparing snapshots, or by monitoring the actions of the user. The description of each change together with any superimposed information about the change will be stored in the archive. Changes will trigger a notification mechanism that will identify related nodes and propagate the change event to all possibly affected information objects.

Multiversion archiving. Automatically archive each new "release" of the data, following a distributed approach for storing information. The archived data will be replicated in several nodes in order to increase efficiency and guarantee the availability and preservation of information.

Longitudinal query capabilities. Answer questions efficiently with complex conditions on the provenance and evolution of information objects. It will be possible to express snapshot queries on previous instances of the data and their relationships, and also pose longitudinal queries that cut across snapshots to give insight about the *how*s and *why*s of the current state of information.

Provenance support. Since in the LOD cloud RDF triples are usually replicated, to assess various forms of data quality, such as trustworthiness, reputation and reliability it is crucial to determine

[20] www.freebase.com
[21] www.geneontology.org,www.biopax.org, www.nlm.nih.gov/research/umls, www.co-ode.org/galen
[22] bio2rdf.org
[23] www.openphacts.org

the origins of published LOD worldwide. This essentially calls for representing and reasoning on the provenance of LOD, as they are transformed by declarative SPARQL queries or inferred through logic programs. Instead of computing each possible annotation, such as trust scores, independently during data sharing, an alternative approach is to record abstract provenance information for capturing the relationship among source and derived data along with the query operators that were involved in the derivations. This provenance information can then be materialized in the repository when the data is imported and used later to

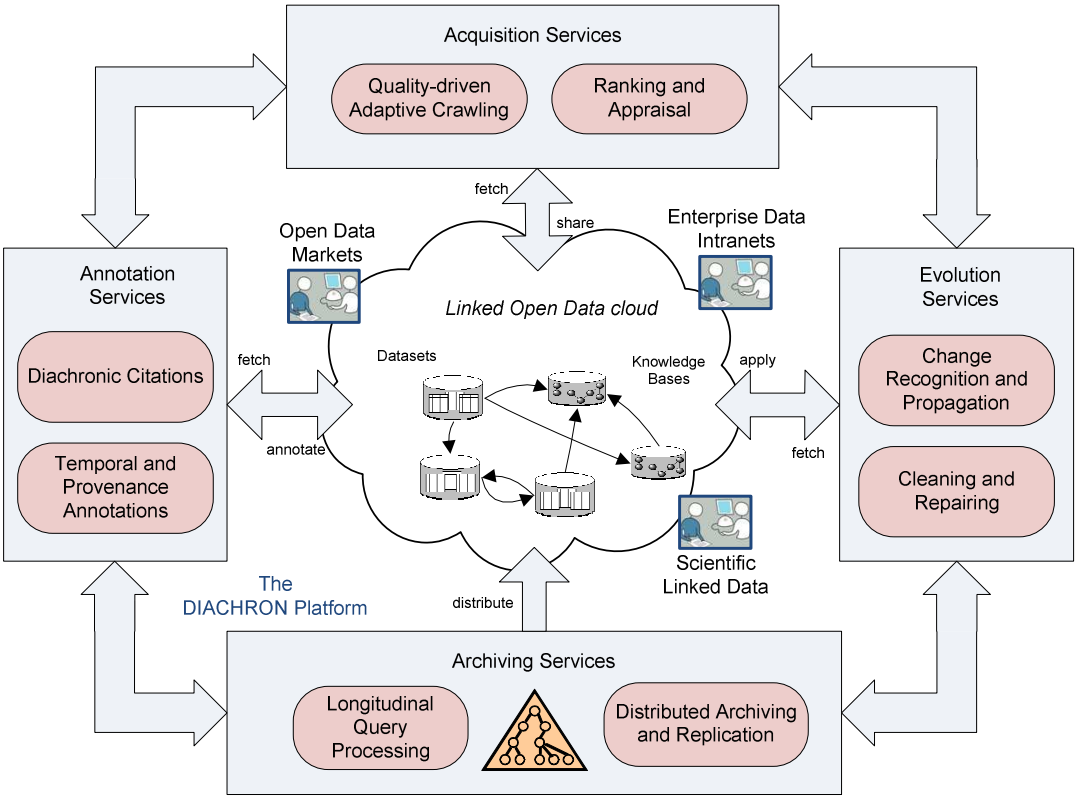

Figure 3: The DIACHRON Platform

compute annotations "on the fly", based on annotations on source data and how they were combined through query operators for a particular application.

Towards this direction, we propose a platform for diachronic linked data, called *DIACHRON*. The platform is not intended to replace existing standards and tools, but rather to complement, integrate, and co-exist with them, by building on previous efforts of the Linked Data community. Figure 3 depicts the overall architecture. Notably, we foresee four groups of services for *long-term LOD accessibility and usability*: *acquisition*, *annotation*, *evolution*, and *archiving* services.

The *acquisition module is* responsible for harvesting LOD datasets published on the Data Web and assessing their quality w.r.t. critical dimensions such as accuracy, completeness, temporal consistency or coverage. It includes services for:

- *Ranking LOD datasets* according to various quality dimensions. Depending on each application scenario, appraisal of LOD datasets in an archive may be based on the spatial or temporal *coverage (data for more geographic regions vs. more updated data versions)* of acquired datasets.
- *Crawling datasets on the Web or Intranets* based on their quality criteria. Rather than fetching locally several datasets and then a posteriori assess their quality, DIACHRON will consider quality conscious-crawling services.

The *annotation module* is responsible for enriching LOD with superimposed information regarding temporal validity and provenance of the acquired datasets. The appropriate *granularity level* of such annotations will be investigated w.r.t. concrete application needs. Depending on the application, this superimposed information is not provided by default by the original LOD providers. Furthermore, unique identifiers of LOD datasets need to be determined in order to enable diachronic LOD citations (i.e., immune to changes) both from printed-material (traditional paper citations) as well as other LOD datasets (data links). It consists of services for:

- *Diachronic citations* based on persistent URIs of LOD datasets. In this way, DIACHRON can track the evolution of LODs monitored in the Data Web, as well as provide "persistent citations" [11], i.e., references to data and their metadata that do not "break" in case those data are modified or removed.
- *Temporal and provenance annotations*. Given that LOD datasets change without any notification while they get freely replicated on the Data Web, understanding where a piece of data (or metadata) came from and why and how has obtained its current form is also crucial for appraisal. To this end, we additionally need to understand the meaning of such piece of data eventually by considering alternative interpretations w.r.t. original production and usage context.

The *evolution* module is responsible for detecting, recording and managing changes of LOD datasets monitored on the Data Web. It provides services for:

- *Cleaning and repairing LOD datasets* based on declarative semantics. These services intend to assist LOD curators in enhancing the quality of the harvested LOD datasets with a minimum possible human intervention. In particular, DIACHRON is interested in coping with LOD inconsistences arising due to newly acquired information (e.g., changes in scientific knowledge), revisions to their intended usage or simply errors (when LOD changes are propagated across the Web from one copy to the other, or even when the employed integrity constraints themselves get revised).
- *Change recognition and propagation* by monitoring and comparing snapshots of LOD datasets. DIACHRON will pay particular attention to the LOD change language used to produce deltas that can be interpreted both by humans and machines. This is a crucial need in various tasks such as the synchronization of autonomously developed LOD versions, or visualizing the evolution history of a particular LOD dataset e.g., during curation and will form a critical input for automated error detection. DIACHRON will finally investigate the possibility for accessible recording of evolving LOD datasets registered in the system.

The *archiving module is* responsible for storing and accessing multiple versions of annotated LOD datasets as presented in the previous modules and services. It comprises services for:

- *Multi-version Archiving* based on internal hierarchical structure for representing LOD datasets that is amenable to compression of inherently redundant information, as well as to query the evolution history of LOD. The archived data will be replicated in several nodes in order to enable community-based preservation of LODs.
- *Longitudinal querying* featuring complex conditions on the recorded provenance and change information of archived LOD datasets. It will be possible to express longitudinal queries that cut across snapshots to give insight about the hows and whys of the current state of information.

The *DIACHRON Platform* will integrate these services into a cohesive framework and will be accessible not only directly from the users, but also from applications that would like to exploit the potential of individual services and components.

5. RELATED WORK

The bulk of the LOD research efforts conducted so far has been focused on effective techniques for publishing data on the Data Web (see LATC[24] project), efficient data management support (see LOD2[25] and PlanetData[26] projects), large scale interlinking and analysis infrastructures (see LATC and CUBIST[27] projects) as well as LOD technology assessment methods (see SEALS[28]) project. The goal of LOD2 is to develop an integrated tool stack [13] for improving the quality of data published on the Web, close the performance gap between relational and RDF data management and establish trust on the Linked Data Web. In DIACHRON, we will re-use some results from LOD2, in particular knowledge base refactoring and repair methods, which play a role in data evolution. The SEALS project is developing a reference infrastructure known as the SEALS Platform to facilitate the formal evaluation of semantic technologies. This allows both large-scale evaluation campaigns to be run (such as the International Evaluation Campaigns for Semantic Technologies) as well as ad-hoc evaluations by individuals or organisations. CUBIST is an EU funded research project with a visionary approach that leverages BI to a new level of precise, meaningful and user-friendly analytics of data by following a best-of-breed approach that combines essential features of Semantic Technologies, Business Intelligence and Visual Analytics. PlanetData is a network of excellence on large scale data management that aims to establish a sustainable European community of researchers that supports organisations in exposing their data in new and useful ways. This is motivated by the increasing reliance of business on *public* data, the use of linked data principles in vertical markets and the increasing volumes of scientific, social and government data. Finally, Linked open data around-the-clock (LATC) aims to support people and organisations to better publish and consume LOD. It offers a *24/7 Interlinking Platform* as a cloud offering to generate RDF links between datasets in the LOD Cloud. LATC offers a library of open source toolkits that cover all stages of the Linked Data publication and consumption process. The LATC project publishes information about Institutions and Bodies in the European Union as LOD to seed the EU data cloud. While DIACHRON will draw on the outcomes of those projects for the implementation of specific parts of its platform, it is worth noting that none of the above projects considers the preservation aspect for LOD. *DIACHRON nicely complements these projects by introducing preservation, evolution, and lifecycle management facilities for the linked data and advances the state-of-the-art by tapping into the problem of preserving entire networks of linked data and knowledge bases.*

[24] atc-project.eu
[25] lod2.eu
[26] www.planet-data.eu/project
[27] www.cubist-project.eu
[28] www.seals-project.eu

6. CONCLUSIONS

In this paper, we argued that a wide range of users and applications would benefit from a framework for managing the preservation of evolving linked data ecosystems. In our view, the temporal aspects should be considered explicitly in the design of algorithms and tools for managing linked data. We presented three representative use cases from the open, enterprise and the scientific data domains, respectively, to demonstrate the real need for evolution and preservation support. We discussed a number of problems we consider as closely related, and proposed a high level architecture of a framework that would tackle those problems.

7. REFERENCES

[1] Reference model for an open archival information system (OAIS). Technical Report CCSDS 650.0-B-1, Consultative Committee on Space Data Systems, 2003. ISO standard 14721:2003, available at: http://nost.gsfc.nasa.gov/wwwclassic/documents/pdf/CCSDS-650.0-B-1.pdf.

[2] C. Bizer, T. Heath, T. Berners-Lee: Linked Data – The Story So Far. Special. Issue on Linked Data, International Journal on Semantic Web and Information Systems (IJSWIS), 2009.

[3] C. Bizer, A. Jentzsch, R. Cyganiak:. 4th State of the Web of Data. In LDOW 2011.

[4] T. Heath, C. Bizer, 2011. Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.

[5] A. Latif, A. Saeed, Patrick Hoefler.The Linked Data Value Chain: A Lightweight Model for Business Engineers. In ISEMANTICS 2009.

[6] J. Umbrich, M. Hausenblas, A. Hogan, A. Polleres, S. Decker: Towards Dataset Dynamics: Change Frequency of Linked Open Data Sources. In LDOW 2010.

[7] L. Zhou L Ding, T. Finin T. How is the Semantic Web evolving? A dynamic social network perspective Computers in Human Behavior (2010), doi:10.1016/j.chb.2010.07.024

[8] A. Ntoulas, J. Cho, C. Olston. What's New on the Web? The Evolution of the Web from a Search Engine Perspective. In WWW 2004.

[9] P. Tsialiamanis, L. Sidirourgos, I. Fundulaki, V. Christophides, P. Boncz: Heuristic based Query Optimisation for SPARQL. In EDBT 2012.

[10] J. Euzenat, P. Shvaiko: Ontology matching. Springer 2007: 1-333.

[11] P. Buneman, G. Silvello: A Rule-Based Citation System for Structured and Evolving Datasets. In IEEE Data Eng. Bull. 33(3): 33-41 (2010).

[12] T. Hey, S. Tansley, K. Tolle (editors). The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research. 2009

[13] S. Auer, L. Bühmann, J. Lehmann, M. Hausenblas, S. Tramp, B. van Nuffelen, P. Mendes, C. Dirschl, R. Isele, H. Williams, O. Erling: Managing the life-cycle of Linked Data with the LOD2 Stack. In Proceedings of International Semantic Web Conference (ISWC'12) 2012.

[14] Mc Kinsey Global Institute: Big data: The next frontier for innovation, competition, and productivity, 2011.

[15] GRDI2020 Consortium. GRDI2020 Final Roadmap Report on Global Research Data Infrastructures: The Big Data Challenges. D4.1, March 2012.