# Schema Extraction for Privacy Preserving Processing of Sensitive Data

Lars Christoph Gleim[1], Md. Rezaul Karim[2,1], Lukas Zimmermann[3], Oliver Kohlbacher[3], Holger Stenzhorn[3], Stefan Decker[1,2], and Oya Beyan[1,2]

[1] Informatik 5, RWTH Aachen University, Aachen, Germany
[2] Fraunhofer FIT, Sankt Augustin, Germany
[3] Chair of Methods in Medical Informatics, University of Tübingen, Germany

**Abstract.** Sharing privacy sensitive data across organizational boundaries is commonly not a viable option due to the legal and ethical restrictions. Regulations such as the EU General Data Protection Rules impose strict requirements concerning the protection of personal data. Therefore new approaches are emerging to utilize data right in their original repositories without giving direct access to third parties, such as the Personal Health Train initiative [16]. Circumventing limitations of previous systems, this paper proposes an automated schema extraction approach compatible with existing Semantic Web-based technologies. The extracted schema enables ad-hoc query formulation against privacy sensitive data sources without requiring data access, and successive execution of that request in a secure enclave under the data provider's control. The developed approach permit us to extract structural information from non-uniformed resources and merge it into a single schema to preserve the privacy of each data source. Initial experiments show that our approach overcomes the reliance of previous approaches on agreeing upon shared schema and encoding a priori in favor of more flexible schema extraction and introspection.

**Keywords:** Semantic Web, Linked Data, RDF, Schema, Privacy, Data Access, Distributed Systems, Query Design, Personal Health Train

## 1 Introduction

Data driven methods play an increasingly important role for cost efficient and timely research results and effective decision support [45] throughout numerous domain such as economics [3], education [42], manufacturing [49], healthcare and life sciences [1, 39, 48].

At the same time, the data that build the foundation of these models oftentimes underlies strict sharing requirements. For example, in the sensitive healthcare domain, although first responders, hospitals, and many other stakeholders already collect valuable data for data-driven research and treatment today, large portions of this data remain inaccessible to the majority of stakeholders – largely due to ethical, administrative, legal and political hurdles that render data sharing infeasible [14].

In practice, this leads to an inability to access large amounts of data crucial for a variety of tasks such as the optimization of decision support systems, first response systems and data-driven research. At the core of this issue lies the lack of an effective mechanism to allow for data access in a legally certain, sustainable and cost-efficient manner without extensive delays.

For example, learning health systems, allowing for data-driven research on sensitive data such as electronic health records (EHRs), have long been said to bear the potential to "fill major knowledge gaps about health care costs, the benefits and risks of drugs and procedures, geographic variations, environmental health influences, the health of special populations, and personalized medicine." [19]. While a variety of such systems have been proposed [18, 19, 22, 26], practical implementation has so far not become a reality, likely due to the aforementioned hurdles.

In order to enable data economy in privacy-sensitive domains and effective reuse of existing data and research, novel approaches are emerging to overcome these limitations. One of those approaches is the Personal Health Train (PHT) framework, which aims to bring algorithms and statistical models to data sources, rather than sharing data with the third parties such as researchers. The main benefit of the PHT approach is its ability of utilizing all the data, including the sensitive and private information, without data having to leave from the original data source. One of the main challenges of the approach is that data users such as researchers are required to develop their models without having a grasp of the actual data. Unless there are universally agreed data set descriptions, there is a need to create and communicate a schema – that is information about the structure of the data – to enable writing queries for heterogeneous data resources.

The key contributions of this paper consist of an automated approach for extracting task-relevant schema from RDF data sources for the formulation of data selection and integration queries without direct access to the data and a corresponding integration with an information system architecture that allows for the subsequent evaluation of that query in a secure enclave.

The rest of the paper is structured as follows: Section 2 describes some related work and the basic foundation. Section 3 formulates the key challenges of schema extraction from sensitive data without sacrificing privacy, followed by the description of our proposed schema extraction approach from existing data in section 4. Additionally, it also demonstrates how to perform the data selection and integration using the extracted schema. Section 5 then outlines initial experimental result based on a sample use case. Finally, some future works have been mentioned before concluding the paper in section 6.

## 2 Related Work

In order to facilitate knowledge discovery for both humans and machines, the FAIR data principles [53] have been proposed: A set of guiding principles to make research and scientific data Findable, Accessible, Interoperable, and Re-usable. These guidance principles promise to help in the discovery, access, integration and analysis of task-appropriate scientific data and associated algorithms and workflows. Thus, FAIR is gaining a lot of attention and increasing adoption.

Core to realizing these principles are Semantic Web technologies [7], which provide a framework for data sharing and reuse by making the semantics of data machine interpretable. Particularly the directed, graph-based data model RDF [10, 36, 40] in conjunction with formal conceptualizations of information models, semantics and encoding conventions in RDF vocabularies and ontologies takes an important role.

As such, RDF Schema (RDFS) [8] and the Web Ontology Language (OWL) [43] provide a proven framework in order to describe (but not necessarily enforce) the structure and semantics of data. Substantially, RDFS introduces the concepts of classes and properties as well as basic relations between them. OWL – a computational logic-based language – extends upon these concepts in order to represent rich and complex knowledge about things, groups of things, and relations between them.

In the context of this work, we use the term 'schema' to refer to the semantic and structural annotation of data using especially these two vocabularies.

On the other hand, the classical notion of schema as the formal definition of the shape that data needs to comply with in order to be valid (i.e. schema validation and enforcement) also exists in the Semantic Web with the Shape Expression Languages (ShEx) [46] and the Shapes Constraint Language (SHACL) [37]. At this time, they are however not part of common data encoding standards, vocabularies or ontologies and as such will not be regarded further in this work.

Nevertheless, using RDFS and OWL, it is possible to create domain-specific, optionally interoperable vocabularies and ontologies, which may declare e.g. term or concept equivalences and dependencies between each other and subsequently enable interoperability across individual encodings.

Popular examples include the Ontology for Biomedical Investigation (OBI) [2] in the biology and healthcare domain, the GoodRelations ontology [32] in eBusiness and the DCAT vocabulary [41], which is used for the general purpose metadata annotation of datasets and data catalogs.

In the context of eHealth systems, first-class support for the Semantic Web is becoming more and more prominent with popular candidates such as HL7 FHIR [6] and SNOMED CT [13] providing corresponding ontologies, as well as the establishment of clear guidelines for dataset descriptions such as the HCLS Community Profile [25].

Various high-quality catalogs of freely reusable vocabularies that provide the description of the data that are available for an easy discovery of suitable ontologies. Examples include the Linked Open Vocabulary (LOV) [51] and the BioPortal [44] project.

Related ideas using schema export and import for federated data access date back to as early as 1985 [31] but it is only recently that the idea has received more attention in the context of the Semantic Web.

Kellou-Menouer and Zoubida [34] propose a schema discovery approach based on hierarchical clustering instead of data annotations thus leading to an approximate schema. Florenzano et al. [21], Weise et al. [52] and Dudáš et al. [15] introduce approaches focused on schema extraction for visualization of the data structure but do not consider publishing or reuse of the extracted schema. Benedetti et al. [4, 5] propose an interesting related approach for schema extraction, visualization and query generation but do not consider interoperability issues and rely on custom mechanisms for schema storage.

## 3 Motivation

Recently, Jochems et al. [33] and Deist et al. [12] introduced two related promising Semantic Web-based approaches in the context of the PHT initiative, founded on the key concept of bringing research to the data rather than bringing data to the research. As such the underlying information system architecture enables learning from privacy sensitive data without the data ever crossing organizational boundaries, maintaining control over the data, preserving data privacy and thereby overcoming legal and ethical issues common to other forms of data exchanges.

The general approach of this underlying system may be outlined as follows:

1. Initially, both the client and data provider agree upon a set of attributes or features, such that all participating data providers have corresponding sources of (privacy sensitive) data.
2. Then each data provider encodes their data using an (also agreed upon) ontology or vocabulary, converting it into RDF representation. This process yields proper Linked Data [30] and thus enables semantic interoperability [11].
3. The resulting RDF data is deployed to a private triple store at each location, providing a private SPARQL [47] query endpoint, which is not directly accessible by the client.
4. A SPARQL data query is then formulated based on the previously agreed upon encoding and a corresponding distributable processing algorithm defined.
5. The shared query is then executed locally at each data provider against their respective triple store and the returned data processed using the corresponding algorithm.
6. The local results are combined into a global one.
7. Depending on the approach, steps 5 and 6 may be further iterated.

While these approaches – introduced in the context of the PHT initiative – work well when multiple parties agree on jointly collecting, encoding and evaluating data in advance – such as is the case for conducting individual coordinated studies – they solve the issue of interoperability by agreeing on a single shared knowledge representation and encoding methodology a priori (steps 1-3 in the above process). In an optimal setting where agreeing on a single shared and global information model and encoding, reuse of diverse and existing data could always be directly accomplished with this approach.

However to our knowledge, so far all corresponding efforts have been unsuccessful. At the time of writing the popular `https://fairsharing.org/` portal indexes 1055 databases using 1136 standards, suggesting that in practice, each collected dataset and domain much rather tends to introduce its own encoding methodology.

Thus when trying to reuse diverse existing data, especially without direct access to the data, ad-hoc data selection and integration facilities (corresponding to the first two steps of the classical Knowledge Discovery in Databases (KDD) process [20]) are indispensable.

For a client without direct access to the data, this process is however typically infeasible since it inherently relies upon inspection of the structure of the data. In this setting, in order to allow for the effective design of such queries (corresponding to step 4 in the above approach), a proper description of the structure of the available data – a schema of the data – is required.

This schema should further be compatible with standard Semantic Web tools for interoperability and thus be available as RDFS and OWL vocabulary via a SPARQL endpoint. While OWL provides a powerful set of modeling primitives, in the context of this work we focus on RDFS and the OWL `owl:equivalentClass`, `owl:equivalentProperty` and `owl:sameAs` predicates, which we deem most relevant in order to enable interoperability and the effective formulation of selection and integration queries.

As a result, the schema not only contains everything that is needed in order to create data queries (i.e. using SPARQL), but also conveys far less privacy critical information than the actual data. As such it can be published publicly without privacy concerns in many scenarios.

In the following we describe an automated approach for schema extraction from RDF data which allows for the formulation of data selection and integration queries without direct access to the data and the subsequent evaluation of that query in a secure enclave.

## 4 Proposed Approach

In this section, we discuss the proposed approach. First, we describe the schema extraction technique. Then we show how the extracted schema can be used further for the data selection and integration.

### 4.1 The schema extraction

We propose an approach for schema extraction based on exploiting the key characteristics of RDF, RDFS, and OWL. RDF data encoded in compliance with aforementioned vocabularies inherently include metadata about their semantics and structural relationships.

For the schema extraction, the `rdf:type` relation plays the key role, as it declares data points to be instances of specific data types or classes. Anything that is a type in the sense of occurring as the target of this relation thus automatically becomes part of the schema. Additionally, any relation (that is any identifier occurring in the predicate position of a subject-predicate-object triple) which occurs in the data is itself a part of the schema and is included as well.

As such, the entire schema of a given RDF data set can be extracted using a single SPARQL CONSTRUCT query as depicted in listing 1.1. For this approach, we assume OWL entailment regime [23] support of the SPARQL endpoint and proper inclusion of the used vocabularies in the triple store.

```
1 CONSTRUCT {?s ?p ?o}
2 WHERE {
3   {[] a ?s}
4   UNION {[] ?s []}.
5  ?s ?p ?o
6 }
```

**Listing 1.1.** SPARQL schema extraction query using entailment regime

As stated earlier, the preceding query constructs an RDF graph (line 1) containing all the directly describing triples `?s ?p ?o` that occur in the tripe store but having only the following subjects:

1. ?s that are used as RDF types (line 3),
2. ?s that are used as predicates (line 4)

According to the SPARQL entailment regime, all the subclass relationships, transitive properties etc. used in the data are automatically resolved and included too. The query however only extracts direct properties and as such some complex constraints such as OWL disjointness axioms are not extracted properly, which we however consider to be irrelevant for the task of query formulation.

Note that we define the relevant subset of all available schema information to be that which is actually used by the data, i.e. the instantiated schema, and thus only extract that.

Since in practice few SPARQL endpoints actually support any kind of entailment, it is alternatively also possible to extract the schema directly using the SPARQL 1.1 Property Paths [38] feature, independent of entailment support on the endpoint. A corresponding SPARQL query is depicted in listing 1.2.

```
1 CONSTRUCT {
2   ?a ?b ?c; a rdfs:Class.
3   ?d ?e ?f; a rdf:Property.
4 } {
5  { SELECT ?a ?b ?c { [] a/(owl:equivalentClass|owl:
      sameAs|rdfs:subClassOf)* ?a OPTIONAL { ?a ?b ?c }
       } }
6  UNION
7  { SELECT ?d ?e ?f { [] ?x []. ?x (owl:
      equivalentProperty|owl:sameAs|rdfs:subPropertyOf)*
       ?d OPTIONAL { ?d ?e ?f } } }
8 }
```

**Listing 1.2.** SPARQL 1.1 schema extraction query using property paths

The query constructs a graph of all RDFS Classes and RDF Properties and their direct properties which are either directly instantiated or used in the dataset, a generalization of one used in the dataset and equivalent resources.

Corresponding to RDF 1.1 Semantics [29] we detect instantiated RDF properties as any IRI used in predicate position (c.f. rdfD2) and annotate them accordingly in line 3 and 7. We detect instantiated classes based on the RDFS axiomatic triple `rdf:type rdfs:range rdfs:Class` as any object of the RDF type predicate in line 2 and annotate them accordingly in line 5.

For both properties and classes, we resolve corresponding generalizations directly using the relevant RDFS entailment patterns (rdfs5, rdfs7, rdfs9, rdfs11) and concept equivalences using OWL's `owl:equivalentClass`, `owl:equivalentProperty` and `owl:sameAs` predicates. While `owl:sameAs` is only supposed to be used for the declaration of equivalence between individuals, it is commonly misused in practice and as such deliberately included in this query.

As such the presented approach is capable of extracting the relevant (i.e. instantiated) schema from a given RDF dataset which can subsequently be used for SPARQL query design without requiring access to the original data.

### 4.2 Data selection and integration using the extracted schema

Once the schema is extracted, the resulting schema can be publicly exposed using a dedicated SPARQL endpoint. It is then possible to use existing SPARQL query writing assistance tools (i.e. query builder) such as OWLPath [50], QueryVOWL [28] or VSB [17] together with the extracted schema for schema introspection aided design of data selection and integration queries. An overview of available tools can be found in [24].

The workflow of the proposed architecture is illustrated in figure 1, which depicts the communication between client and data provider over a public network. In this scenario, the data provider's internal communication within its private network is highlighted by the bounding box. In preparation for client usage, the

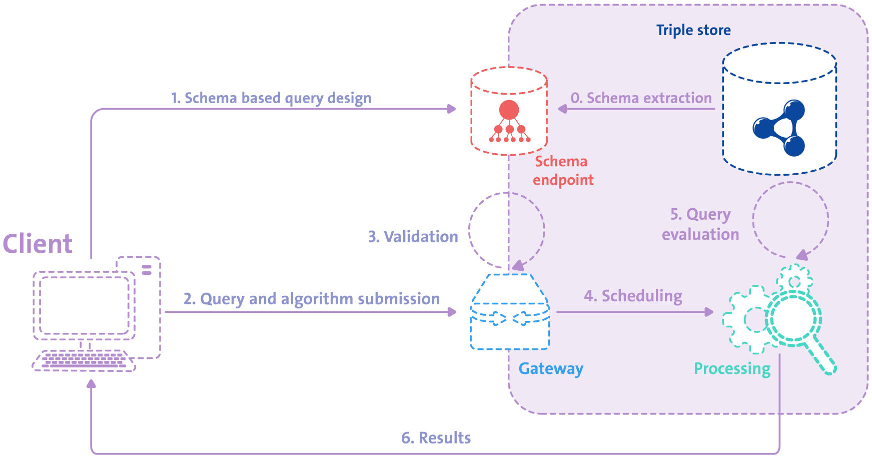

**Fig. 1.** Workflow of the proposed architecture

schema of the data stored in the private `triple store` is extracted in *step 0* using the approach presented above and deployed to a publicly accessible `schema endpoint`.

The `client` can now start to create a SPARQL query in *step 1*, using a query builder of their choice in conjunction with the `schema endpoint` for introspection. The query is then sent to a `submission endpoint` acting as the gateway between the data provider and the client in *step 2*. For the scope of this work, We assume that this requests includes algorithmic means of data anonymization, ensuring its results are no longer privacy sensitive and that validation is done manually.

Once validated, the request is scheduled in *step 4* for processing within a secure enclave (`processing`), where the query and algorithm are evaluated (*step 5*). This is analogous to the approach proposed by Jochems et al. [33] and Deist et al. [12] as detailed in section 2. Finally, only the processing result is returned to the client in `step 6` without ever directly granting access to the data.

## 5 Experimental Results

In order to allow for a first evaluation of the proposed approach, a simple test case was constructed. A number of records containing personal information of individuals such as name, birthday and phone number were constructed, encoded using the foaf and schema.org vocabularies and deployed to a private triple store. The relevant schema subset was then extracted using the query depicted in listing 1.2 and its correctness and completeness validated manually. Subsequently the public schema was used in order to create a data selection and integration query, which was successfully executed against the private data endpoint.

In order to evaluate the effectiveness of the schema extraction process, we employ the HCLS core statical measures to compare the characteristics of the full schema.org [27] and foaf [9] vocabularies, their union and the extracted schema. The results, depicted in table 1, show that the number of extracted triples is significantly lower compared to the original count and roughly 3.3 percent of the union of both vocabularies. As intended, only the subset of the vocabularies that actually describes the private dataset is extracted, allowing for focused query design based on only the relevant schema, thus saving cognitive as well as computational effort during schema introspection.

**Table 1.** HCLS core statistics of the full schema.org [27] and foaf [9] vocabularies, their union and the schema extracted using our proposed approach.

| **HCLS** | **number of** | **schema** | **foaf** | **both** | **extracted** |
|---|---|---:|---:|---:|---:|
| 6.6.1.1 | triples | 8427 | 631 | 9058 | 317 |
| 6.6.1.2 | unique, typed entities | 1617 | 84 | 1701 | 86 |
| 6.6.1.3 | unique subjects | 1619 | 86 | 1705 | 86 |
| 6.6.1.4 | unique properties | 15 | 15 | 23 | 14 |
| 6.6.1.5 | unique objects | 476 | 38 | 508 | 43 |
| 6.6.1.6 | unique classes | 31 | 9 | 38 | 8 |
| 6.6.1.7 | unique literals | 3193 | 154 | 3335 | 105 |

Since the extracted schema also contains explicit equivalence information (for example between the `foaf:Person` and `schema:Person`, which is in this case only declared in the schema.org vocabulary) it is possible to explicitly design queries considering the corresponding implications at query design time without relying upon inference support of the SPARQL endpoints. As such it may provide an additional building block for enabling efficient interoperability across different data codings.

## 6 Conclusion and Outlook

In this paper, we proposed an automated way of schema extraction from Linked Data in RDF format. Our proposed approach enables the introspection supported development of SPARQL queries without access to the actual data. We presented a system architecture to realize the overall workflow of the approach. From the users perspective, our approach enables in query formulation against privacy sensitive data sources and successive evaluation of that request in a secure enclave at the data provider's end.

With this architecture, we can overcome the reliance of previous approaches on agreeing upon shared schema and encoding a priori in favor of more flexible schema extraction and introspection.

This method promises to provide a key building block in enabling efficient reuse of data across a variety of domains. In conjunction with advanced distributed learning and processing systems, the approach could be used in order to overcome existing data sharing hurdles and unlock hidden value in existing data silos.

In the future, we plan to extend this work by exploring integrations with query federation engines and access control, such as the SAFE query federation engine [35]. We also plan to provide a more extensive performance analysis and evaluation in order to show the effectiveness of this approach.

## References

1. Abernethy, A.P., Etheredge, L.M., Ganz, P.A., Wallace, P., German, R.R., Neti, C., Bach, P.B., Murphy, S.B.: Rapid-learning system for cancer care. Journal of Clinical Oncology 28(27), 4268–4274 (2010), https://doi.org/10.1200/JCO.2010.28.5478, pMID: 20585094
2. Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M.H., Bug, B., Chibucos, M.C., Clancy, K., Courtot, M., Derom, D., Dumontier, M., et al.: The ontology for biomedical investigations. PloS one 11(4), e0154556 (2016)
3. Basole, R.C., Russell, M.G., Huhtamäki, J., Rubens, N., Still, K., Park, H.: Understanding business ecosystem dynamics: a data-driven approach. ACM Transactions on Management Information Systems (TMIS) 6(2), 6 (2015)
4. Benedetti, F., Bergamaschi, S., Po, L.: Online index extraction from linked open data sources. In: LD4IE@ ISWC. pp. 9–20 (2014)
5. Benedetti, F., Bergamaschi, S., Po, L.: Visual querying lod sources with lodex. In: Proceedings of the 8th International Conference on Knowledge Capture. p. 12. ACM (2015)
6. Beredimas, N., Kilintzis, V., Chouvarda, I., Maglaveras, N.: A reusable ontology for primitive and complex hl7 fhir data types. In: Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE. pp. 2547–2550. IEEE (2015)
7. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific american 284(5), 34–43 (2001)
8. Brickley, D.: Rdf vocabulary description language 1.0: Rdf schema. http://www.w3. org/TR/rdf-schema/ (2004)
9. Brickley, D., Miller, L.: Foaf vocabulary specification 0.91 (2007)
10. Cyganiak, R., Wood, D., Lanthaler, M., Klyne, G., Carroll, J.J., McBride, B.: Rdf 1.1 concepts and abstract syntax. W3C recommendation 25(02) (2014)
11. Decker, S., Melnik, S., Van Harmelen, F., Fensel, D., Klein, M., Broekstra, J., Erdmann, M., Horrocks, I.: The semantic web: The roles of xml and rdf. IEEE Internet computing 4(5), 63–73 (2000)
12. Deist, T.M., Jochems, A., van Soest, J., Nalbantov, G., Oberije, C., Walsh, S., Eble, M., Bulens, P., Coucke, P., Dries, W., Dekker, A., Lambin, P.: Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT. Clinical and Translational Radiation Oncology 4, 24–31 (2017), http://linkinghub.elsevier.com/retrieve/pii/S2405630816300271
13. Donnelly, K.: Snomed-ct: The advanced terminology and coding system for ehealth. Studies in health technology and informatics 121, 279 (2006)

14. Doshi, P., Jefferson, T., Del Mar, C.: The imperative to share clinical study reports: recommendations from the tamiflu experience. PLoS medicine 9(4), e1001201 (2012)
15. Dudáš, M., Svátek, V., Mynarz, J.: Dataset summary visualization with lodsight. In: International Semantic Web Conference. pp. 36–40. Springer (2015)
16. Dutch Tech Center For Life Sciences: Manifesto of the Personal Health Train consortium (2017), `https://www.dtls.nl/wp-content/uploads/2017/12/PHT_Manifesto.pdf`
17. Eipert, L.: Metadatenextraktion und vorschlagssysteme im visual sparql builder. INFORMATIK 2015 (2015)
18. Embi, P.J., Payne, P.R.: Clinical research informatics: challenges, opportunities and definition for an emerging domain. Journal of the American Medical Informatics Association 16(3), 316–327 (2009)
19. Etheredge, L.M.: A rapid-learning health system. Health affairs 26(2), w107–w118 (2007)
20. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI magazine 17(3), 37 (1996)
21. Florenzano, F., Parra, D., Reutter, J.L., Venegas, F.: A visual aide for understanding endpoint data. Visualization and Interaction for Ontologies and Linked Data (VOILA! 2016) p. 102 (2016)
22. Friedman, C.P., Wong, A.K., Blumenthal, D.: Achieving a nationwide learning health system. Science translational medicine 2(57), 57cm29–57cm29 (2010)
23. Glimm, C.O.B.: Sparql 1.1 entailment regimes (2012)
24. Grafkin, P., Mironov, M., Fellmann, M., Lantow, B., Sandkuhl, K., Smirnov, A.V.: Sparql query builders: Overview and comparison. In: BIR Workshops (2016)
25. Gray, A.J., Baran, J., Marshall, M.S., Dumontier, M.: Dataset descriptions: Hcls community profile. Interest group note, W3C (May 2015) http://www. w3. org/TR/hcls-dataset (2015)
26. Greene, S.M., Reid, R.J., Larson, E.B.: Implementing the learning health system: from concept to action. Annals of internal medicine 157(3), 207–210 (2012)
27. Guha, R.V., Brickley, D., Macbeth, S.: Schema. org: evolution of structured data on the web. Communications of the ACM 59(2), 44–51 (2016)
28. Haag, F., Lohmann, S., Siek, S., Ertl, T.: Queryvowl: Visual composition of sparql queries. In: International Semantic Web Conference. pp. 62–66. Springer (2015)
29. Hayes, P.J., Patel-Schneider, P.F.: Rdf 1.1 semantics. w3c recommendation, february 2014. World Wide Web Consortium. Retrieved from https://www.w3.org/TR/2014/REC-rdf11-mt-20140225 (2014)
30. Heath, T., Bizer, C.: Linked data: Evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology 1(1), 1–136 (2011)
31. Heimbigner, D., McLeod, D.: A federated architecture for information management. ACM Transactions on Information Systems (TOIS) 3(3), 253–278 (1985)
32. Hepp, M.: Goodrelations: An ontology for describing products and services offers on the web. In: International Conference on Knowledge Engineering and Knowledge Management. pp. 329–346. Springer (2008)
33. Jochems, A., Deist, T.M., van Soest, J., Eble, M., Bulens, P., Coucke, P., Dries, W., Lambin, P., Dekker, A.: Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept. Radiotherapy and Oncology 121(3), 459–467 (2016), `http://dx.doi.org/10.1016/j.radonc.2016.10.002`
34. Kellou-Menouer, K., Kedad, Z.: Schema discovery in rdf data sources. In: International Conference on Conceptual Modeling. pp. 481–495. Springer (2015)

35. Khan, Y., Saleem, M., Mehdi, M., Hogan, A., Mehmood, Q., Rebholz-Schuhmann, D., Sahay, R.: Safe: Sparql federation over rdf data cubes with access control. Journal of biomedical semantics 8(1), 5 (2017)
36. Klyne, G., Carroll, J.J.: Resource description framework (rdf): Concepts and abstract syntax (2006)
37. Knublauch, H., Ryman, A.: Shapes constraint language (shacl). W3C Candidate Recommendation 11, 8 (2017)
38. Kostylev, E.V., Reutter, J.L., Romero, M., Vrgoč, D.: Sparql with property paths. In: International Semantic Web Conference. pp. 3–18. Springer (2015)
39. Lambin, P., Rios-Velazquez, E., Leijenaar, R., Carvalho, S., van Stiphout, R.G., Granton, P., Zegers, C.M., Gillies, R., Boellard, R., Dekker, A., et al.: Radiomics: extracting more information from medical images using advanced feature analysis. European journal of cancer 48(4), 441–446 (2012)
40. Lassila, O., Swick, R.R.: Resource description framework (rdf) model and syntax specification (1999)
41. Maali, F., Erickson, J., Archer, P.: Data catalog vocabulary (dcat). W3C Recommendation 16 (2014)
42. Marsh, J.A., Pane, J.F., Hamilton, L.S.: Making sense of data-driven decision making in education (2006)
43. McGuinness, D.L., Van Harmelen, F., et al.: Owl web ontology language overview. W3C recommendation 10(10), 2004 (2004)
44. Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D.L., Storey, M.A., Chute, C.G., et al.: Bioportal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research 37(suppl_2), W170–W173 (2009)
45. Power, D.J., Sharda, R., Burstein, F.: Decision support systems. Wiley Online Library (2015)
46. Prud'hommeaux, E., Labra Gayo, J.E., Solbrig, H.: Shape expressions: an rdf validation and transformation language. In: Proceedings of the 10th International Conference on Semantic Systems. pp. 32–40. ACM (2014)
47. Prud'hommeaux, E., Seaborne, A., et al.: Sparql query language for rdf (2006)
48. Shiboski, S., Shiboski, C., Criswell, L., Baer, A., Challacombe, S., Lanfranchi, H., Schiødt, M., Umehara, H., Vivino, F., Zhao, Y., et al.: American college of rheumatology classification criteria for sjögren's syndrome: A data-driven, expert consensus approach in the sjögren's international collaborative clinical alliance cohort. Arthritis care & research 64(4), 475–487 (2012)
49. Simchi-Levi, D.: Om forum—om research: From problem-driven to data-driven research. Manufacturing & Service Operations Management 16(1), 1–22 (2017)
50. Valencia-García, R., García-Sánchez, F., Castellanos-Nieves, D., et al.: Owlpath: An owl ontology-guided query editor. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 41(1), 121–136 (2011)
51. Vandenbussche, P.Y., Atemezing, G.A., Poveda-Villalón, M., Vatant, B.: Linked open vocabularies (lov): a gateway to reusable semantic vocabularies on the web. Semantic Web 8(3), 437–452 (2017)
52. Weise, M., Lohmann, S., Haag, F.: Ld-vowl: Extracting and visualizing schema information for linked data. Visualization and Interaction for Ontologies and Linked Data (VOILA! 2016) p. 120 (2016)
53. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al.: The fair guiding principles for scientific data management and stewardship. Scientific data 3 (2016)