LingHub: An Adapted LOD Cloud Repository for Linguistic Data

Bettina Klimek, Martin Brümmer, Natanael Arndt, Sebastian Hellmann, Thomas Riechert, and Henri Knochenhauer

University of Leipzig, Institute of Computer Science, AKSW Group,
Augustusplatz 10, D-04109 Leipzig, Germany
{lastname}@informatik.uni-leipzig.de,
http://aksw.org

Abstract. In this paper we present LingHub - a metadata repository that accounts for the needs of data providers and users from the linguistic community. Focussing at the diversity of currently existing data repository models we found that none of these are universally applicable. Instead, large Linked Open Data repositories have been developed next to closed high quality data repositories restricted to a selected user community. To tackle the issue for the linguistic domain, this paper points out specific requirements for storing and reusing open and unrestricted datasets while maintaining a valuable provenance chain. We provide an overview of already existing tools and software components and explain how these were used to set up the adaptable metadata repository, LingHub, that fills the identified gaps.

Keywords: data repositories, open data, provenance, metadata, best practices and guidelines

1 Motivation

In 2011, the European Commission published its Open Data Strategy[1] defining the following six barriers for "open public data": (1) A lack of information that certain data actually exists and is available, (2) A lack of clarity of which public authority holds the data, (3) A lack of clarity about the terms of re-use, (4) Data which is made available only in formats that are difficult or expensive to use, (5) Complicated licensing procedures or prohibitive fees, (6) Exclusive re-use agreements with one commercial actor or re-use restricted to a government-owned company.

Taking these barriers as a starting point, it becomes obvious that no universal remedy exists. Instead these high-level problems have to be broken down into smaller challenges for specific use cases and domains in order to provide adapted solutions. In this submission, we will refine these challenges to create a blueprint for the implementation of a metadata repository called *LingHub*.

[1] http://europa.eu/rapid/press-release_MEMO-11-891_en.htm

LingHub intends to collect metadata for datasets adhering to the intersection of the following four topics:

- language resources and linguistic data;
- linked data (including the precursor data sources);
- openly licensed or freely available;
- scientific data value chains (with an emphasis on provenance).

We furthermore argue that data from the **linguistic domain** on the one hand adds additional challenges due to the inherent complexity of the data, while on the other hand it is of special importance for the Semantic Web in general. Linguistic data has the unique ability to aid developers in bridging the gap between user-intelligible messages and machine-readable data, also called Human-computer interaction by providing dictionaries, gazetteers, annotated text corpora, term hierarchies, wordnets with lexical information, translations for training machine translation systems. Normally for English, all these resources are readily available and easy to find (even with an open license), for less-spoken languages, however, such as !Kung[2] or Tagalog such resources are rare in the first place and nearly impossible to find, if they are not correctly indexed and their metadata described properly. **Achieving complete metadata with high quality at low cost** requires a repository with an intelligent, user-friendly workflow that generates metadata where appropriate, while allowing crowd-source co-evolution to ease maintenance. We designed *LingHub* as a specialization of Datahub(cf. Sect. 5) to cater to this and other requirements elaborated below. The main use case, which lead us to create an OntoWiki-based [2] repository, was the fact that no existing repository or data model was able to properly describe and curate metadata for DBpedia, which has developed into a large-scale, multilingual knowledge base[3], which is often the only available Linked Data source for less-spoken language.

Our work originates from the discussions within the Working Group for Open Data in Linguistics (OWLG)[5]. OWLG is one of 19 working groups of the Open Knowledge Foundation[4] (OKFN), each promoting open knowledge in their specific domain. OWLG mainly exists of individual researchers organized in a grassroots movement without institutional orchestration. The realization of LingHub, however, is supported by the recently started LIDER EU project[5] whose goal is to support communities such as OWLG and also provide: (1) A set of guidelines and best practices for building and exploiting LOD-based resources in multimedia and multilingual content analytics and for developing NLP services on top of Linguistic Linked Data. (2) A reference architecture for Linguistic Linked Data built on top of existing and future platforms and freely available resources. (3) A long-term road map for the use of Linked Data for multilingual and multimedia content analytics in enterprises.

[2] ~14k-18k speakers according to `http://en.wikipedia.org/wiki/!Kung_language`

[3] accepted at SWJ: `http://www.semantic-web-journal.net/content/dbpedia-large-scale-multilingual-knowledge-base-extracted-wikipedia-0`

[4] OKFN working groups: `http://okfn.org/wg/`

[5] LIDER EU project: `http://lider-project.eu`

2 General Problem Statement

Although a large variety of metadata repositories and data catalogues exist for extensive and heterogeneous types of data, there is a basic framework every metadata repository is bound to: the term data. Taking it back to its etymological origins reveals that data is the plural form of datum meaning 'given'. The verb "give" has a valency of three taking a subject and two objects: someone gives something to somebody. In consequence data is bound to a trinity consisting of (1) the producer or source of the data, (2) the entity that constitutes the data itself and (3) the receiver or user of the data [7]. This trinity supplies important details for implementation, when considering the three different perspectives of metadata collection:

1. The data producer (also publisher or provider) is in a unique position to provide contextual information about the circumstances under which the data came into existence (be it derived or original) such as license, provenance, contributors or additional notes;
2. The data itself can be inspected for metadata, gaining insight about technical information such as file size, format, up-time, availability, validity and in case of RDF and Linked Data: used ontologies, links to other datasets, class and property structure;
3. Finally, the user of the data can supply valuable input on the required information by her use case and the usefulness of data categories. Users are able to give direct feedback about the quality of data and metadata. Users also have a direct incentive to correct and curate data, so that either their applications work properly or they are able to answer research questions (e.g. how many open dataset are available?).

3 Specific Challenges and Requirements

Complete Metadata vs Entry Barrier: Acquiring the desired metadata is challenging, because it can increase the effort of publishing considerably. The more metadata the publisher is required to supply, the less he might be willing to do the effort. Although the effort of metadata entry can be reduced by automatic extraction, some information may only be known to the producer of the data. The trade-off between completeness of metadata and high publishing effort and entry barriers has to be carefully considered.

Incentives for Metadata Curation: For data providers it is already a great effort to publish data properly. Even more so, if they have to add information to many different metadata repositories and propagate updates manually as well. The risk of such statically entered metadata is that it becomes obsolete. As much metadata as possible should be kept alongside the dataset as instead, to guarantee up-to-date, authorative information without increasing the workload for the publisher. By constant reloading, this approach also allows to re-generate and track extensive technical information like SPARQL endpoint URL and up-time, data format and validity and ontology usage. In addition to automatic

extraction, distributing the load of maintenance among users to complete and consolidate data in a wiki approach is critical.

Metadata Coverage vs Metadata Quality: Crowd-sourced data often has better coverage, but is not of the same high quality as expert-curated sources. Provenance has therefore to be extended to metadata entries, allowing crowd-based judgement of data and metadata quality by users and approved domain experts.

License and Attribution: Licensing is a constant issue in Linked Open Data. A compromise is needed between the needs of data publishers to have their data used correctly and their work cited accordingly and the wishes of users to access and reuse the data as freely as possible. Granular licensing and attribution information has to be explicitly included in the metadata to assure publishers to choose open instead of closed licenses. At the same time, data that is Open Access but closed license should also be included to increase repository scope and allow publishers to retain more rights if needed.

Metadata Visualisation: Metadata presentation is widely reduced to textual data in the form of "field name - field value" tuples. However, complex metadata like links between datasets and dataset categories can more effectively be represented in diagrams and images, giving the user better tools to find relevant data and understand its relations.

Granularity: Current repositories only account for dataset-level metadata. However, datasets may include resources that are derived from different sources, which constitutes the need for intra-dataset provenance. RDF can be used to point into datasets and add granular provenance information if it is not provided in the original data. Too granular metadata, however, is cumbersome to maintain and may stop to be useful.

Lack of best practices and clear guidelines: Finally, we are lacking best practices and clear guidelines for modelling and publishing provenance and other metadata. Extent and desired granularity are as unclear, as the way of explicating it and conveying it to the user. There is no clear remedy for these challenges, rather, design decisions have to be grounded on the needs of producers and users alike.

4 The LingHub Metadata Model, Re-used Vocabularies

The Dublin Core Metadata Terms (DCTERMS)[6] have widely been adopted for basic metadata annotation. Properties like `dcterms:title`, `dcterms:creator`, `dcterms:created` and `dcterms:source` are used in many datasets as a minimal core of dataset description. There are, however, metadata scopes unique to the Semantic Web, that cannot be described by DCTERMS.

The Vocabulary of Interlinked Datasets (VoID)[7] tackles these points by specifically defining a vocabulary for dataset description. It defines a class `void :Dataset`, properties particular to RDF datasets like `void:triples`, `void:example`

[6] DCTERMS: `http://purl.org/dc/terms/`
[7] VoID: `http://www.w3.org/TR/void`

`Resource` and `void:dataDump` as well as criteria on how to use other vocabularies, one of them being DCTERMS. The LingHub datamodel uses VoID extensively as seen in Figure 4. The central class of the LingHub model is `void:Dataset`. Most of its properties can be automatically extracted from the dataset itself, should it also use VoID. Other properties, like `dcterms:issued` that describes the date the dataset was published on or other statistical information, like `void:triples`, are computed by LingHub without user input from the dataset itself. Every dataset has to interlink at least one resource of the class `ling:Category` that serves to categorize the data in a domain specific way. Categories are defined by the domain experts and can be extended and edited by users. Consensus is reached by discussion either in LingHub directly or on the related mailing lists [8]. The `void:LinkSet`, that normally contains links from one dataset to another, was replaced by a simple `ling:links` property.

To further adopt the metadata model to the linguistic domain, a special focus was laid on the granular description of provenance information. The need for this arises out of the nature of the linguistic datum as such, since language can be represented on various levels ranging from phonetic transcription of an audio file to the whole semantic description of a certain language. As a result, the majority of linguistic data originates from another linguistic data source. The derived data then emerges by editing the source data [7] and often differs in respect to the goals and theorems underlying the original dataset compilation. In order to make scientific use of any linguistic data it is therefore essential for a researcher to have as much provenance information as possible on the dataset used within their own research areas. Neither VoID nor DCTERMS by themselves are able to express provenance information beyond single persons or related sources of the data.

The Provenance Ontology (Prov-O)[9] with its classes `Entity`, `Agent` and `Activity` including their respective properties is used to capture a complete, fine-grained provenance chain. Entities serve as a super-class of datasets and are all kinds of sources from which datasets are derived that are not datasets themselves, like books or other primary sources. Agents are all kinds of persons involved with dataset creation or maintenance. Activities are mediators in the derivation, integration or aggregation of datasets from other datasets and sources, denoting in their description what changes were made in the process. Slight changes were made to the properties that link Activities and Entities to the Agents. The provenance ontology defines the properties `prov:wasAssociatedWith` and `prov:wasAttributedTo` to link Activities and Entities to the respective Agents. This way, the important "creator" relation, for example, would have to be modeled as an Activity that has a `dcterms:description` literal denoting "created the dataset" and that is linked to the creator via `prov:wasAssociatedWith`. Although this is a valid model in its generality, it obscures the finer-grained semantics of the most basic Entity-Agent relations in literal strings of arbitrary format. Therefore, LingHub uses DCTERMS properties to link datasets to persons, creator, publisher and contributor.

[8] `http://lists.okfn.org/pipermail/open-linguistics/2013-October`
[9] Prov-O: `http://www.w3.org/ns/prov#`

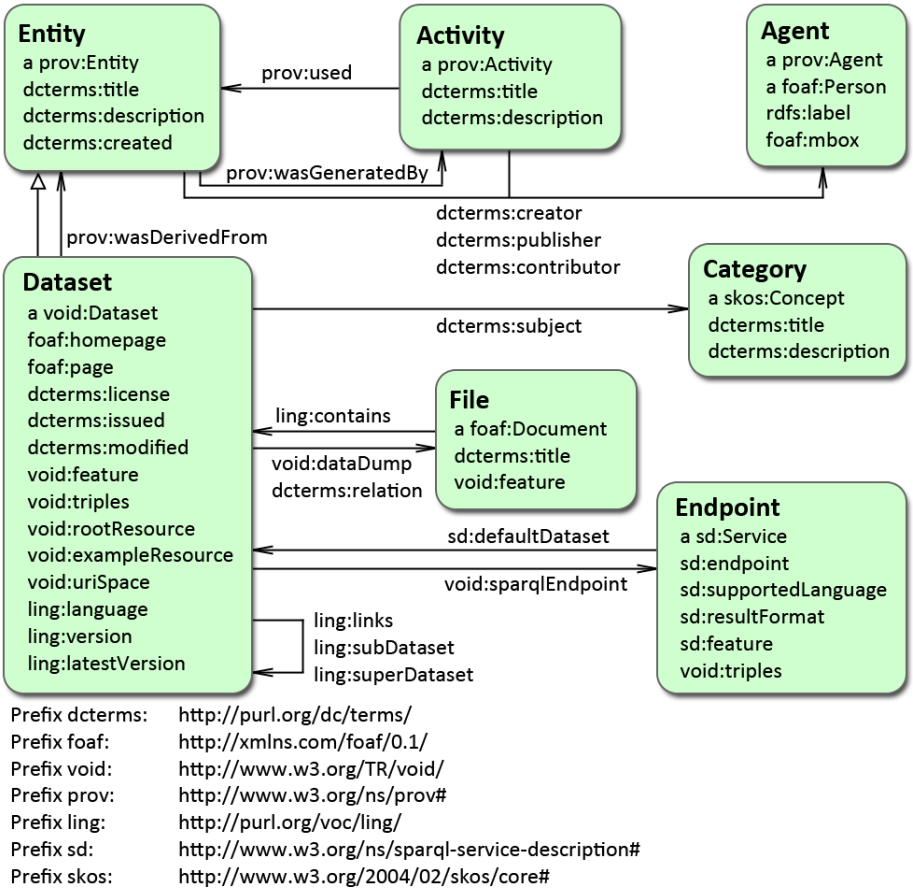

Fig. 1. LingHub metadata model combining VoID and Prov-O

However, there are still use-cases that could not be sufficiently modeled by these vocabularies. Especially, DBpedia [8] is not one monolithic dataset, but an aggregation of 119 different language DBpedia datasets[10] as well as a repository to collect links[11]. Although modelling this relation with the Provenance Ontology was possible[12], this obscures the semantic relation between the dataset and its sub datasets. In LingHub, this is expressed by the `ling:subDataset` and `ling:superDataset` relations. Currently it is also impossible to describe the single language DBpedia dump file's formats sufficiently, because they are compressed files (gzipped) containing RDF serializations. Their MIME type, for example, can only express the "outer" data format, not the format of the contained data. In

[10] DBpedia dataset: `http://downloads.dbpedia.org/3.9/`

[11] `https://github.com/dbpedia/dbpedia-links`

[12] e.g. using a `prov:Activity` that describes the aggregation and `prov:used` the different language DBpedia datasets

the LingHub model, there is a specific class for additional files, that can either be specifically denoted as dataset dumps or just be of arbitrary nature, like diagrams or further documentation. This best practice enables the annotation of the inner format with the `void:feature` property and the Unique URIs for File Formats[13].

Finally, SPARQL-endpoint metadata will be generated automatically via SPARQL Endpoints Status [4][14] and described with properties from the SPARQL 1.1 Service Description[15].

5 Existing Data Repositories (Related Work)

The LingHub Metadata Model arose out of the aim to build a repository that accounts for the needs of data providers and users from the linguistic community. Even though various language resource repositories already exist, we found that none of them conforms to the necessary requirements outlined so far. Before a detailed discussion of this is carried out, a brief presentation summarizes several prominent and well-established data repositories in the following section.

Datahub[16] is a platform developed by OKFN that enables users to upload, group and search open data. It is built on the Comprehensive Knowledge Archive Network (CKAN) software and provides metadata about (Linked Data) datasets. Datahub hosts data unrestricted to any domain and enables both data providers and users to edit dataset entries. By intention the metadata provided for most of the datasets is flat and simple. On the one hand, this facilitates the ease of upload and addition of datasets for providers but on the other hand it reduces the usefulness of the metadata and the accessibility of the data itself. However, the most significant problem of the dataset management at datahub is the absence of detailed provenance information. Merely, an activity stream documents who applied changes in the dataset entry. Datasets in RDF format for example do not reference the data source they were derived from in a proper way. As a consequence Datahub is incapable of adequately describing datasets with multiple source datasets and files such as the well-known DBpedia.

META-SHARE. Another data repository is presented by META-SHARE[17] which is part of the *Multilingual Europe Technology Alliance (META)* and aims at providing quality language resources. Focusing on the processing of the metadata a strictly provider-driven account is taken. META-SHARE assumes a high quality of the data it hosts, because only scientific institutions are allowed to add datasets. Once the data is integrated into the repository no further data validation is conducted. For the data providers, however, it is often infeasible to update the metadata in regular intervals and there is no issue reporting by user requests. This contributes to a rather statical way of data storage and leads to

[13] `http://www.w3.org/ns/formats/`
[14] SPARQL Endpoints Status: `http://sparqles.okfn.org/`
[15] `http://www.w3.org/TR/sparql11-service-description/`
[16] DATAHUB: `http://datahub.io`
[17] META-SHARE: `http://www.meta-share.eu`

an unbalanced data repository favouring data preservation but - given the realm and possibilities of the Semantic Web and Linked Data - contributing little to effective data reuse.

Language Resources Evaluation Map (LRE Map). A quick and structured access to information on language resources is provided by the Language Resources Evaluation Map[18]. It originated 2010 at the LREC conference where all contributing authors were asked to fill in a form asking for information about the language resources they used. In the following years, authors from other conferences joined this procedure as well, so that a matrix of nearly 4,000 language resources emerged. A more specific search is realized through various metadata values that can be multiply selected. However, the LRE Map does not host any of the resources it lists in the catalogue and provenance only goes as far as to state the project page of the dataset (not the download links). Therefore the LRE map basically represents a collection of metadata that is restricted to a selected group of data providers and offers only a display of language resource names and categories to the user.

Common Language Resources and Technology Infrastructure (CLARIN). A complex and sustainable repository network for digital language data is achieved by the Common Language Resources and Technology Infrastructure, that is shared by selected research centers in the humanities and social sciences across different countries. The realization of metadata compilation and storage is based on gathering metadata descriptions which are used to set up a so called Component Metadata Instance (CDMI) that creates a CDMI metadata file for each language resource. Metadata categories are fixed and bound to ISOcat categories but editable by every CLARIN member. Therewith greater dynamics within the metadata maintenance is assured. Furthermore, datasets can be downloaded and run on private computers or used online with certain offered visualisation tools. These useful and time-saving features of accurately documented datasets, however, are neither in Linked Data format nor openly accessible to everyone.

Linguistic Data Consortium (LDC). One of the most specific repositories for language resources is maintained by the Linguistic Data Consortium[19] (LDC) which is run by the University of Pennsylvania. Among textual resources like corpora and lexical language sources other valuable language materials such as audio and video files are supplied as well. Anyone who compiled a linguistic dataset is able to publish her resources using a corpus submission form[20]. Each publication proposal will be checked for completeness and errors in collaboration with LDC staff members. That way, the LDC assures a high quality of all datasets provided. On the downside, it has to be mentioned that Linked Open Data formats are neither required nor supported by the LDC. The contrary is the case: all datasets are bound to closed licenses and subject to fees for both LDC members paying less and for non-members charged full prices. As a consequence the high quality of the provided datasets is repressed by the commercial data

[18] LRE Map: `http://wwww.resourcebook.eu`
[19] LDC: `http://ldc.upenn.edu`
[20] `http://ldc.upenn.edu/data-management/providing/submission`

supply. Within the context of the Semantic Web this only adds to the obstacles that need to be overcome in order to make qualitative data freely accessible to everyone.

6 Comparison of Data Repository Features

The presentation of the five resource repositories above displays the diversity of the implementations used. Taking the trinity of data (cf. section 2) as the basic framework for data repository setup, a detailed comparison of repository features is applied to the mentioned repositories as well as to the new LingHub repository (cf. Table 1).

The most basic feature concerns the *repository content*. A user visiting a website for the first time primarily wants to know what data is offered and if it corresponds to the data she is looking for. All repositories with the exception of Datahub identify themselves as domain specific regarding the linguistic domain. With the certainty that the desired data is stored the user will go on to search more specificaly for datasets. Therefore an understanding of the structure of the repository content is necessary. The more complex and confusing the repository design is the sooner a user will get frustrated by the search procedure and is likely to leave the repository unsatisfied. Due to its small size, an overview of LDC is very easy. LRE Map and LingHub both offer free-text search and facetted-browsing, which decrease the effort to understand how data can be found. Since Datahub contains large amounts of datasets a quick read through the repository description is necessary to find linguistic data specifically, Datahub offers structuring via groups and tags. META-SHARE, however, being part of the more complex META-NET project requires more examination of the repository internal data organisation. The most opaque repository structure is exhibited by CLARIN, because the repository content is visible to registered members only and the membership involves a three step authorization process in addition to the complex repository structure. Once the structure is understood all six repositories allow for a domain-specific search via fixed linguistic categories.

As soon as the user has chosen a dataset, the second feature of *accessibility* becomes crucial. Ideally every dataset provides Linked Open Data or is at least licensed to be free, open and reusable for everyone. But this only applies to LingHub and Datahub. The other repositories supply open and closed datasets (META-SHARE, LRE Map) or closed datasets only (CLARIN, LDC). Even if datasets are open in terms of licenses all repositories with the exception of LingHub and Datahub diminish the accessibility of these through various membership restrictions.

Assuming that the user has successfully gained access to the dataset of interest and used it already for her research, questions of *data contribution* arise next. If she found mistakes in the metadata and would like to correct them, only LingHub and Datahub allow for a direct editing of the metadata entry. The four remaining repositories reserve editing rights for the data providers or repository

	Datahub	**META-SHARE**	**LRE Map**	**CLARIN**	**LDC**	**LingHub**
Repository Content						
Domain specific repository	✗	✓	✓	✓	✓	✓
Effort to Understand Repository Structure	quick reading up on repository description necessary	intense reading up on repository description necessary	low	high	low	low
Ease of Dataset Access	easy, open access - search box on main page	easy - 2 steps through the network	easy - search box on main page	difficult - restricted 3 step authorization process	easy - language resources on website navigation	easy - search box on main page
Custom Visualizations (via API)	✓	✗	✗	✓	✗	✓
Data Categorization	editable by everyone (groups, tags)	fixed categories according to compiled metadata	fixed categories	fixed categories via ISOcat	fixed categories	editable by everyone (OWL and SKOS)
Data Validation	✗	✗	✗	✗	✗	✓
Accessibility						
User Registration	free - required for data providers only	restricted - required for data providers only	restricted - required for data providers and users	restricted - for countries, institutions only, registration takes 2 days	restricted - universities, foundations, organisations only, with membership fee	free - for data providers and users
Openness of Data	LOD	free and closed licenses	free, redirect to data source	closed data only	closed data only - purchaseable for nonmembers	LOD, programatic download
Dataset Hosting	✓	✓	✗	✓	✓	✗
Data Contribution						
Openness of Dataset Entry	everyone	members only	members only	members only	everyone, but controlled by LDC before upload	everyone
Openness of Metadata Entry	everyone	members only	members only	members only	members and providers only	everyone
User Voting on Metadata	✗	✗	✗	✗	✗	cf. Sect. 8
Bulk Editing	✗	✗	✗	✗	✗	via SPARUL and automatic generation of metadata
Provenance Constitution						
Kind of Provenance	source URL	information on dataset creation	source URL	information on dataset creation	information on dataset creation, category, source	data provenance chain
Derived Data Specification	✗	✗	✗	✗	✗	✓
Provenance Support	rudimentary (source homepage only)	no provenance	rudimentary (source homepage only)	mandatory provenance	fine-grained mandatory provenance	mandatory (Prov-O)
Metadata Processing						
Metadata Maintainance	manually (crowd-sourced by all stakeholders)	manually, only data provider	manually, only LREC community	manually, only CLARIN partners	manually, only LDC members	automatically and manually (crowd-sourced by all stakeholders)
Automatic Generation of Metadata (link analysis)	✗	✗	✗	✗	✗	✓
Granularity of Metadata	per dataset	1 to 5 per dataset	16 core elements per dataset	15 core elements per dataset	7 elements per dataset	7 core elements per dataset

Table 1. Comparison of selected language resource repositories according to their features.

members. Equally an unrestricted integration of datasets into the repository has been arranged in LingHub and Datahub but not in META-SHARE, LRE Map and CLARIN. The latter ones restrict the upload and editing of datasets to registered members. LDC gives everyone the possibility to submit own datasets, however, strictly controls and adjusts it to LDC standards. Leaving the editing of metadata and dataset information to data providers or authorized persons only leads to a static dataset description that might contain unnoticed mistakes and withholds the community crowd from taking over maintenance tasks.

A central problem within the domain of linguistic data deals with the issue of *provenance constitution*. With linguistics being an empirical research area, data usage demands for an explicit provenance chain. This includes information stating the source as well as the origin of derived datasets. A great variety of provenance information is observable: Datahub and LRE Map merely state the web page URL, forcing the user to collect all necessary provenance information from there. META-SHARE and LDC leave the user with an indication of the data sources and creators. A statement on how CLARIN is treating data provenance is not possible here, because we have no access to the repositories. In none of these five repositories provenance information is provided in as much detail as thorough linguistic research requests. Therefore, LingHub catches up on this deficiency by explicating the appropriate provenance information (cf. Section 4). With the application of a provenance ontology LingHub is even able to specify the source information of derived datasets. Existing RDF language resources are not RDF-native, but often based on primary linguistic data. Not providing a complete provenance chain will render the metadata useless for linguistic research due to the lack of traceability.

The last repository feature discusses *metadata processing* and affects data users and providers alike. The repositories differ in the number of metadata information between 1 to 16 core elements. Thereby META-SHARE compiles more metadata than is actually displayed with the dataset entry, leaving the creation date of the dataset to be the sole explicit metadata information for a part of the datasets. With the exception of LingHub and Datahub, metadata maintenance is done manually by the data providers or authorized members in every repository. The consequences of manual curation of technical information are often inaccurate and not up to date metadata due to the lack of resources. In order to provide sustainable dataset information, LingHub tackles this problem by implementing an automatic generation of metadata via dataset inpection and link analysis (cf. Section 7) next to manually edited metadata. That way the accuracy of the technical metadata is not only assured but also facilitates the effort of dataset upload for the data providers.

After having gained an elaborate insight into the various data repository constructions, three different kinds of data repositories can be identified: (1) User-Centred (Datahub), (2) Provider-Centred (META-SHARE, LRE Map, CLARIN), (3) Data-centred (LDC).

According to the trinity of data framework, the presented repositories - except LingHub - reveal an unbalanced emphasis on only one data component, which

results in the deficiencies outlined so far. As a solution, LingHub instantiates a balanced data repository by encompassing the trinity of data. That means being easy and freely accessible to users, supporting data submission for data providers through an automatic metadata retrieval and assuring high quality datasets via dynamic repository structures cared for by the whole linguistic community.

7 LingHub Components and Implementation

A domain-adapted metadata repository has to face and solve the challenges as described in the sections above if it is set to be a relevant and useful alternative to existing solutions. To tackle the presented issues we have decided to create a modular and extensible system which makes highly use of several components from the LOD2 stack [1][21]. The interaction of those components is depicted in Figure 2. As core of the implementation the OntoWiki [2][22] application framework is used. It serves as a basis to represent the metadata to the user in various, domain-adaptable ways via its site-extension and is amenable to editing via its collaborative wiki functionality. The effort of metadata curation, coverage and quality judgment can be distributed between the users and the producers. LODStats [6][23] is used to analyze the provided data and extract technical and statistical metadata automatically such as number of triples, links to other datasets and ontologies used. Setting guidelines for metadata annotation within the datasets (e.g. providing a data descriptor turtle file), dataset-level metadata such as author, maintainer and provenance can also be automatically extracted, decreasing the overall manual effort and providing better maintainability. Further technical data can be retrieved by using the SPARQL Endpoint Status tool[24] [4] to monitor SPARQL endpoint URLs and up-times. The whole infrastructure is backed by a Virtuoso Open Source powered triple store.

OntoWiki Application Framework. OntoWiki [2] is a collaborative tool for community driven knowledge engineering. Following the idea of wiki-systems, it allows for simultaneous editing of semantic content with an intuitive inline editing mode for RDF triples, similar to WYSIWIG for text documents. It also provides different views on instance data by offering semantic enhanced search strategies and different possibilities to browse the data. It fosters social collaboration aspects by keeping track of changes and allowing comments and discussion on every single part of a knowledge base.

The OntoWiki Site-Extension adds another presentation layer on top of the collaboration platform. It provides a template-based system for publishing the RDF-resource held in the OntoWiki to a non-technical audience. This is used for the realization of the user interface components: dataset import form, presentation and publication of the resources.

[21] LOD2 stack: `http://stack.lod2.eu/`
[22] OntoWiki: `http://ontowiki.net`
[23] LODStats: `http://stats.lod2.eu/`
[24] SPARQL Endpoints Status: `http://sparqles.okfn.org/`

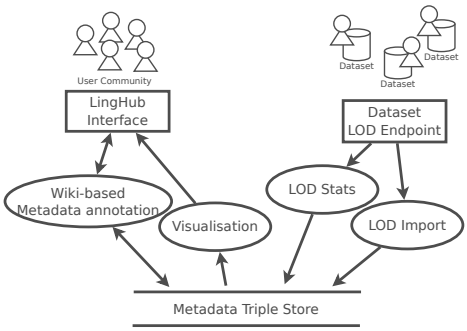

Fig. 2. Data flow diagram of the communication taking place in LingHub between the several components and the public LingHub interfaces.

All these techniques are applied with the ultimate goal of decreasing the entrance barrier for projects and domain experts to collaborate using semantic technologies. Due to the general-purpose approach and its high flexibility, OntoWiki can be seen and used as an application framework for easily building semantic web applications such as LingHub.

OntoWiki itself has no possibility for persistent dara storing and relies on a database backend. Virtuoso is the most mature backend for OntoWiki. The basic version is free, but there is the option for professional support as an enterprise-grade data server. Virtuoso already brings many features such as a SPARQL endpoint, WebID and multi-model support and is accessed by OntoWiki through its ODBC interface.

LODStats Statistic Data Gathering. When dealing with different datasets it is often difficult to obtain a clear picture of the characteristics of the available datasets, like structure, coverage and coherence of the data. LODStats [6] is a statement-stream-based approach for gathering comprehensive statistics about datasets adhering to the Resource Description Framework (RDF). LODStats is based on the declarative description of statistical dataset characteristics. Its main advantages over other approaches are a smaller memory footprint and significantly better performance and scalability. LODStats has been already integrated with the CKAN dataset metadata registry in one way, i.e. fetching all LOD datasets for analysis. We are integrating LODStats, as well in the other direction into our workflow, i.e. to generate statistical metadata and publish them alongside the publisher- and community-authored metadata.

Interfaces and Data Gathering Process

As already depicted in Figure 2 the LingHub repository has two interfaces. One for retrieving metadata about registered datasets and a second one for adding/registering new datasets. The first interface for retrieving the metadata is designed according to the Linked Data principles [3]. It provides an HTML representation, meant to be displayed to a technically unexperienced audience, and an interface to request the RDF resources and to query it with SPARQL. Furthermore, registered users can access the OntoWiki editor view to insert new

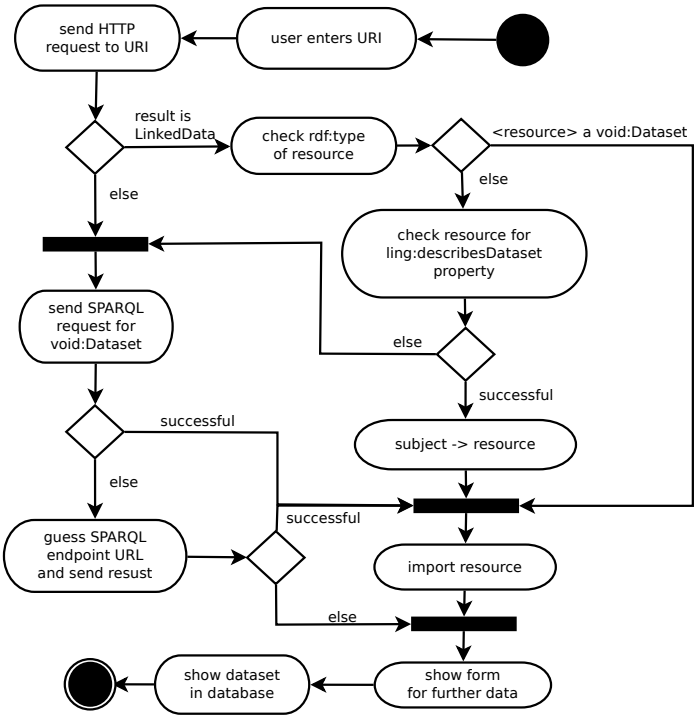

Fig. 3. Activity diagram of the datagathering process after a users submitted a dataset URI to the system.

metadata or suggest changes for already given metadata. The second interface for registering new datasets provides a form to manually add a new dataset URI and optionally additional static metadata. The process of registering a new dataset is depicted in Figure 3. In addition, a Semantic Pingback service [9] provides the possibility to automatically register new datasets. This action can also be used to trigger an update of already registered datasets. After a successful import of the resource a LODStats data gathering process is initiated to generate the technical/statistical metadata and add it to the dataset description.

8 Discussion and Outlook

LingHub is highly relying on crowd-based metadata authoring and cross-linking for linked open data sets. Previous approaches that are only based on meta-information provided by the maintainer of the data sets run the risk of storing outdated and incomplete information. The presented crowd-based co-evolution strategy allows users of datasets to add the description from their specific perspective. Differences by the users of the approach lead to a discussion of the data and thus to a higher accuracy. LingHub is deployed at `http://lgd.aksw.org`

/linghub. As a proof of concept we have added the DBpedia dataset metadata, which can be compared to the entry at Datahub http://datahub.io/dataset/dbpedia. The metadata of DBpedia is loaded from the LingHub model dataset descriptor provided at http://dbpedia.org and e.g. here [25].

The OntoWiki Application Framework already provides a huge list of extensions[26] especially for voting on the quality of resources. Although this feature is crucial to provide reputation-based trust and rating-provenance its realization is yet difficult and was moved to future work. Main problems are for example, whether star-ratings stay valid when major facts of the dataset metadata change. The most promising approach is currently MediaWiki's Sighted Version plugin[27], but effectiveness for metadata needs to be further evaluated.

Acknowledgements. We especially would like to thank our colleagues Christian Chiarcos, John McCrae, Sebastian Nordhoff, Hugh Paterson III and other members of OWLG community and also our fellows from the AKSW research group for their helpful comments and inspiring discussions. This work was supported by the EU projects: LOD2 (GA 257943) and LIDER (GA 610782)

References

1. S. Auer, L. Bühmann, C. Dirschl, O. Erling, M. Hausenblas, R. Isele, J. Lehmann, M. Martin, P. N. Mendes, B. van Nuffelen, C. Stadler, S. Tramp, and H. Williams. Managing the life-cycle of linked data with the lod2 stack. In *ISWC*, 2012.
2. S. Auer, S. Dietzold, J. Lehmann, and T. Riechert. OntoWiki: A tool for social, semantic collaboration. In *CKC@WWW*, 2007.
3. T. Berners-Lee. Linked Data. Design issues, W3C, June 2009. http://www.w3.org/DesignIssues/LinkedData.html.
4. C. Buil-Aranda, A. Hogan, J. Umbrich, and P.-Y. Vandenbussche. Sparql web-querying infrastructure: Ready for action? In *ISWC*. 2013.
5. C. Chiarcos, S. Hellmann, S. Nordhoff, S. Moran, R. Littauer, J. Eckle-Kohler, I. Gurevych, S. Hartmann, M. Matuschek, and C. M. Meyer. The open linguistics working group. In *LREC*, 2012.
6. J. Demter, S. Auer, M. Martin, and J. Lehmann. Lodstats – an extensible framework for high-performance dataset analytics. In *EKAW*. Springer, 2012.
7. C. Lehmann. Data in linguistics. *The Linguistic Review*, 21:175–210, 2004.
8. J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, J. C. Sahnwaldt, C. Stadler, P. van Kleef, S. Auer, C. Bizer, and K. Idehen. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web Journal*, to appear in 2014. http://www.semantic-web-journal.net/content/dbpedia-large-scale-multilingual-knowledge-base-extracted-wikipedia-0.
9. S. Tramp, P. Frischmuth, T. Ermilov, and S. Auer. Weaving a Social Data Web with Semantic Pingback. In *EKAW*, 2010.

[25] https://github.com/dbpedia/dbpedia-links/blob/master/datasets/dbpedia.org/lobid.org/manifestation/metadata.ttl

[26] https://github.com/AKSW/OntoWiki/tree/develop/extensions

[27] http://governancexborders.com/2013/01/29/sighted-versions-in-wikipedia-a-case-of-algorithmic-governance/