data

MDPI

Article

CRC806-KB: A Semantic MediaWiki Based Collaborative Knowledge Base for an Interdisciplinary Research Project

Christian Willmes [1,*], Finn Viehberg [2], Sarah Esteban Lopez [2] and Georg Bareth [1]

[1] Institute of Geography, Department of Geosciences, University of Cologne, 50923 Cologne, Germany; g.bareth@uni-koeln.de

[2] Institute of Geology and Mineralogy, Department of Geosciences, University of Cologne, 50923 Cologne, Germany; finn.viehberg@uni-koeln.de (F.V.); s.estebanlopez@uni-koeln.de (S.E.L.)

* Correspondence: c.willmes@uni-koeln.de

Received: 31 August 2018; Accepted: 20 October 2018; Published: 25 October 2018

Abstract: In the frame of an interdisciplinary research project that is concerned with data from heterogeneous domains, such as archaeology, cultural sciences, and the geosciences, a web-based Knowledge Base system was developed to facilitate and improve research collaboration between the project participants. The presented system is based on a Wiki that was enhanced with a semantic extension, which enables to store and query structured data within the Wiki. Using an additional open source tool for Schema–Driven Development of the data model, and the structure of the Knowledge Base, improved the collaborative data model development process, as well as semi-automation of data imports and updates. The paper presents the system architecture, as well as some example applications of a collaborative Wiki based Knowledge Base infrastructure.

Keywords: data base; information management; knowledge base; metadata; linked data; interoperability; semantic wiki; collaborative web infrastructure; virtual research environment

1. Introduction

This study presents an approach for developing a collaborative research database in the context of an interdisciplinary and inter-institutional research project, the German Research Foundation (DFG) funded Collaborative Research Centre 806[1] (CRC 806). The CRC 806 theme "Our way to Europe", concerns "Culture-Environment Interaction and Human Mobility in the Late Quaternary", and focuses on three major research themes [1]: (i) the climatic, environmental and cultural context; (ii) secondary occurrences of expansion and retreat; and (iii) population changes, mobility and migration in coupled cultural and environmental systems. The project exists since 2009 and was funded in three four-year terms until 2021.

The CRC 806 operates a data management project, that maintains a data management infrastructure named the CRC 806 Database[2] [2–5]. This web accessible frontend of the CRC 806 data management infrastructure implements the data management policy and demands of the CRC 806 project funder, the DFG [6–8]. The CRC 806 Database consists of (i) a data archive and publication platform (CRC806-DB) [3]; (ii) a spatial data infrastructure (CRC806-SDI) [9]; and (iii) a literature and publication database [2] containing all publications produced by the CRC 806, as well as further

[1] http://www.sfb806.de
[2] http://crc806db.uni-koeln.de

features, like a directory for research sites and field campaigns within the project. The here presented Knowledge Base (KB) system was primarily designed to facilitate interactive collaborative research directly in the sense of a truly collaborative web platform, like a Virtual Research Environment (VRE) [10] or a Cyberinfrastructure [11] for data look-up, discovery, data integration, and data analysis, as a project internal environment for sharing and creating in-progress data collections. This project internal KB is called CRC806-KB and is described in this paper in detail.

The research within the CRC 806 has a truly interdisciplinary research setting, and the main research questions within the project are of spatiotemporal context and concern heterogeneous data sources. This entails, that most of the research questions asked can be answered by analyzing spatiotemporal patterns in the given data. Consequently, this led us to built an application that allows these kind of queries on the heterogeneous data of the project.

A circumstance that makes the endeavor to create an integrated data base for the CRC 806 ambitious, is the heterogeneity of the data domains of discourse. And of course, the heterogeneity within the domains and its sub-domains. We deal with data ranging from geoscientific sampling, like core data or sediment and soil analyses, to archaeological site descriptions including dated artefacts and analyses of excavation profiles, to published literature and further publicly available external data of interest for the spatiotemporal context of concern.

The spatial annotation of an archaeological or geoscientific artefact is sufficiently clear, in the case of temporal annotation it is much less clear. And if we look at the integration layer of cultural or environmental classifications, nomenclature and annotations, we find our self in mere chaos. Thus, the development of an integrated data model can almost always be seen as the seek for the smallest valid denominator. Thus, we need a simple to use, collaboratively editable, preferably web-based application to allow the project participants to collect and edit data in a central infrastructure, and provide the possibility to alter and extend the content, structure and data model of the data collection. We found that wikis deliver most of these demanded functionality for editing content in an intuitive web-based collaborative platform. And because we looked for the ability to structure and query the collected content as data, we found Mediawiki[3] with its extension Semantic Mediawiki[4], that allows to store, edit and query structured information in the Wiki, as a perfect fit for our use cases.

2. Related Work, Software and Technology

2.1. Related Work

Data, information and knowledge are closely related terms, but each has its own role in relation to the other. This relation is formalized in a concept called the *Knowledge Pyramid* [12] (see Figure 1), which represents the structural and functional relationship between data, information, and knowledge. Some models of the *Knowledge Pyramid* are extended to include the concept of wisdom above the concept of knowledge, in this case it is called the *DIKW Pyramid*, for *Data*, *Information*, *Knowledge*, and *Wisdom*.

The inference from Figure 1 is that data begets information begets knowledge begets wisdom. An additional inference is that there is more data than information, more information than knowledge, and more knowledge than wisdom [13]. A different formalization of that same concept would be to express these relations in summations, of the form:

$$I = \sum(D)$$

$$K = \sum(I) = \sum\sum(D)$$

[3] https://www.mediawiki.org
[4] https://www.semantic-mediawiki.org

$$W = \sum(K) = \sum\sum(I) = \sum\sum\sum(D)$$

With: W = Wisdom, K = Knowledge, I = Information, D = Data.

Figure 1. The wisdom hierarchy [14], or the DIKW pyramid [13,15] based on the Knowledge Pyramid [12].

One important purpose of the *Knowledge Pyramid* concept is to reflect that the level of abstraction increases from data upwards to wisdom. Thus, the concept puts the relationship between data, information and knowledge into a hierarchical arrangement based on the level of abstraction [2].

The presented approach for creating an internal KB for the CRC 806 was first formulated in [16], at this time called a "bottom-up" approach for data model development and data integration. In this first instance of the approach, the data was imported into a triple store by mapping the data to a prototyped development ontology using python scripts [16]. This approach proved to be complex and hard to develop and maintain, because applications that could make use of the data via SPARQL needed to be developed and caused an additional layer of work. The current instance of this prototype bottom-up approach integrates the data into a semantic Wiki, that delivers sufficient visualization, data input and query interface to the underlying database out of the box. In case of the here facilitated Semantic Mediawiki software [17], it also has the possibility of querying the data through a SPARQL endpoint and exporting in RDF format.

The presented system is a KB, because it aims to combine a lot of information and data, as well as to model a certain knowledge on top of a given data collection. An example of the kind of knowledge created by, and maintained in the system, is the ability of the system to show what research was conducted according a certain theme, an archaeological setting, a time and/or a location. A key focus are tools and interfaces to query and discover spatiotemporal patterns, that were not directly visible from the data collection beforehand. This knowledge is modeled in form of queries on the structured data collection. Thus, the knowledge is represented through the queries on the data collection, and its result. The term KB is justified by the fact that this system resembles a database containing information about resources (databases and datasets), which on their own already contain information about their data.

2.2. Literature Review

Using the terms *knowledge*, *knowledge management* and *knowledge base* in the given context, as introduced in the previous section, these terms need to be contextualized with, and related to the wider research context within the information technology, knowledge engineering and semantic web research domains. Furthermore, some examples of similar applications in the domain of archaeology are given.

According to [18], *knowledge* may be viewed or defined from several perspectives, a state of mind, an object, a process, a condition of having access to information, or a capability. *Knowledge management* is of complex and multi-faceted nature. For example, knowledge may be tacit or explicit; it can refer to an object, a cognitive state, or a capability; it may reside in individuals, groups, documents, processes, policies, physical settings, or repositories. This leads to the insight, that no single or optimum knowledge management or knowledge management system (KMS) can be developed or implemented [18–20]. Ref. [21] defines KMS as a combination of knowledge management practices (KM-Practices), as a set of methods and techniques to facilitate KM development, and knowledge management tools (KM-Tools), as specific systems supporting KM-Practices [22]. Semantic wikis including Semantic Mediawiki [23], as applied for the example implementation presented in this study are KM-Tools according to [17,19,22,24]. In this regard, the here applied KM-Practise consists mainly of the data integration process into the knowledge base system [16], as well as in the formulation of complex queries that are possible on this integrated data base, which reveal new information and knowledge, that was not available before the data was integrated. Similar knowledge base approaches in the archaeology domain are for example [25,26]. These two examples also embrace the here described advantages of semantic wikis of being formal and flexible at the same time. Further interesting insights into the specifics of knowledge work in the archaeology domain in context with data, information, knowledge and digital technologies in general are discussed by [27].

Knowledge management is closely related to the Semantic Web [28] research domain with a focus on the formalization of data models and schemas by application of ontologies [29] and ontology engineering [30]. Those ontologies or data models are highly formalized and developed using specialized software and standards. Ontologies can facilitate data integration of two or more discrete data sets, without the need to map or align the schemas or terminology of the two or more data sets during the integration process. This advantage facilitates the possibility to query on these data sets without the need of additional data integration work. As said, this is not the case for the here presented system, a part of the knowledge created through the CRC806-KB originates from this semi-automated data integration process (see Sections 2.4 and 3.1).

The CRC806-KB does not built on, or implement a formal ontology, in the narrow sense as for example designed by [31] or for example implemented by [32,33]. The CRC806-KB approach refines the internal structure, ontology, or schema on a requirement basis *on-the-fly* [2], for example when integrating a new dataset, the model, structure, schema or ontology can be adapted, as described in Section 2.4, if for example a definition of a term is missing, new data is mapped to the already existing terminology, structure or ontology, as far as possible [2]. The advantage of this approach is its flexibility, because it can be adapted to any domain in a straight forward manner, but its trade-off is the lack of interoperability. To achieve this, a mapping to well defined schema or data model, for example schema.org [34] or Wikidata [35], or any well defined metadata model would be necessary.

2.3. Software and Tools

Collaborative knowledge management is often facilitated through wikis [36]. The Wikipedia project is the most prominent example of a public collaborative wiki. A main disadvantage of conventional wikis was the collaborative editing of structured data and information. Basic wikis, such as MediaWiki without additional extensions, do not have sophisticated mechanisms to reuse or even query for structured information in their content documents, apart from free text search. The usual way of structuring information in wikis is through categorization/tagging and maintenance of documents, pages, lists and directories. Semantic Wikis solve this shortcoming by adding functionality to allow creation, editing and querying of *structured data* in wiki platforms [24].

2.3.1. Mediawiki

The semantic wiki is based on the MediaWiki (MW) [37] wiki implementation. MW is well known as the open source software basis of the famous Wikipedia online encyclopedia, with millions of

content entries and also millions of daily users. The MW software is free and Open Source Software, and implemented in PHP and MySQL. The MW project has a large developer community, on the one hand several full time developers, funded by large cooperations using the software. On the other hand small to mid-size consultancies offering MW based services and also, to a significant amount, the Wikimedia foundation itself, and some of its local chapters (i.e., the Germany chapter, with currently six full time developers for several MW based software projects like WikiData and SMW, as of late 2017). Because of this professional developer community, the MediaWiki software is stable and mature, and facilitated in many business, R&D, educational, NGO and governmental installations. MW is in use in thousands of Wikis around the world, it is almost certainly the world's most popular Wiki software [38].

2.3.2. Semantic Mediawiki

MW is brilliant in facilitating Wiki functionality, like collaborative editing of unstructured text, but it lacks functionality for managing structured information. This is where SMW [23] can help out, it adds the possibility to collaboratively enter and edit structured information in MW [17]. It defines a framework for storing data in a Wiki, and querying it, which has the effect of turning a Wiki into a collaboratively editable database [38].

SMW is a free, open-source extension to MW, that enables to store and query data within the wiki's pages, and offers a full-fledged framework, in combination with many spin-off extensions, that can turn a MW instance into a powerful and flexible Knowledge Management System (KMS). All data created within SMW can easily be published via semantic web formats and data models, allowing other systems to use this data [23].

The SMW extension allows to enter structured semantic data on Wiki pages. This data can then be queried, using the SMW query language *ASK*, through several interfaces within the Wiki and the Mediawiki API as well as an SPARQL endpoint for access from external applications. Query results can be exported in several well known formats, such as CSV, XML, JSON, and more (see Figure 2). It is also possible to display query results directly in the Wiki, using a number of provided so called *Semantic Result Formats*, like tables, data graphs or the Semantic maps Extension for displaying query results on interactive maps.

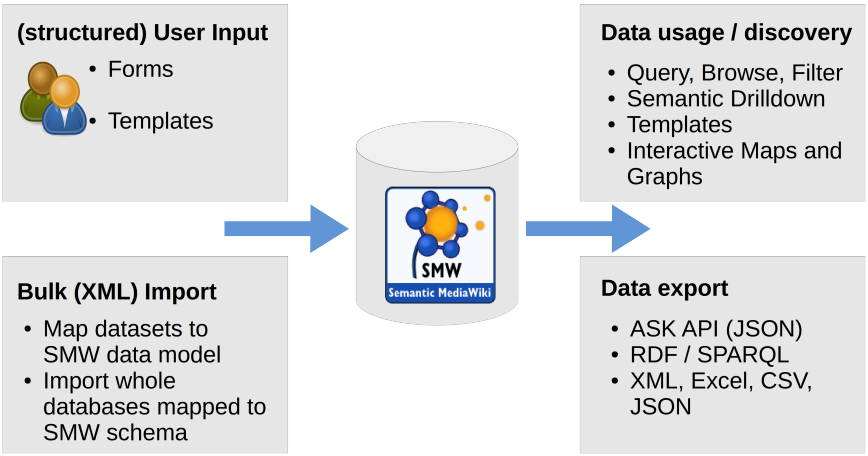

Figure 2. Semantic Mediawiki interfaces. Source: [2].

SMW is applied in related applications and thematic domains. One example for a related application would be the use of SMW for developing an archaeological collaborative research database by [26]. Further interesting applications are for example Collaborative Process Development [39],

Semantic Portal [40], Collaborative Framework for Representation and Harmonization of Clinical Study Data Elements [41], and the application in health sciences [42].

2.4. Data Model Development

Developing and maintaining a data model in SMW can become very complex and cumbersome, because every change of the model has to be applied in several pages (e.g., property pages, template pages, form pages) of the Wiki. To improve the management of this complexity, the Mobo [43] toolkit was introduced at the SMWCon fall 2014 [44] in Vienna. Mobo is a toolset that helps to build SMW structure in an automated Software Design Description (a simplified Model Driven Engineering (MDE)) approach [45]. Schema Driven Development uses annotated data schemas, that specify the expected data structures, as models to generate system artifacts (code, documentation, tests, etc.) automatically [45]. The model is written and developed in YAML or JSON, using the object oriented JSON Schema [46]. The Schema Driven Development approach simplifies the SMW data model development significantly, because the building blocks are more generic, and thus more simple to reuse. The model can therefore be very "DRY" (Don't Repeat Yourself) [45] without unnecessary redundancy (which is normally a well know problem of hitherto prevalent SMW data model administration and development). The target wiki must have the Semantic MediaWiki [23] and Semantic Forms [47] extension installed. It is highly recommended to install the *ParserFunctions* Extension, since mobo's default templates make use of it. But it is possible to adjust/use templates that work without it instead [45].

The main feature of Mobo is the simplified and improved model development workflow (see Figure 3). Semantic MediaWikis can be developed rapidly and modular, leading to a more agile development process. Mobo can run in an interactive mode, automatically validating and uploading the development model in real-time [45].

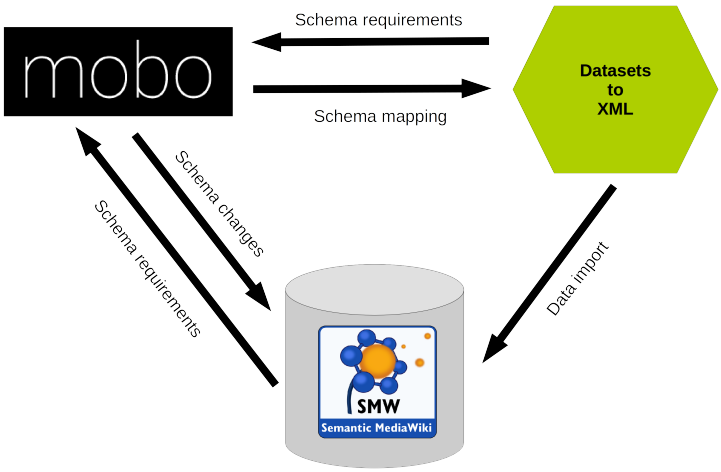

Figure 3. Mobo based data model development. Source: [2].

A further useful feature of Mobo is the possiblitiy to validate the model for its syntax, structure and semantics is also valuable. The syntax is validated in the parsing step of the schema files, by the JSON and YAML libraries, that Mobo includes. The structure is checked by validating the model against the meta-schema. Additionally it is possible to check for semantic errors by implementing according logics [45]. Further external JSON or YAML linters (Tools that flag suspicious usage in software written in any computer language) and validators can be facilitated to evaluate the model. Since MediaWiki wikitext markup misses validation capabilities, the ability of the generator to validate the development model is a big benefit in comparison to using wikitext directly [45]. Another benefit

of formulating the SMW model in JSON Schema or YAML, is the possibility to use collaborative software development methods and tools like Git. Furthermore there is the possibility to version the development stages of a model and also to work collaboratively on the model schema.

3. System Architecture and Implementation

Developing an integrated research database for a large interdisciplinary research project is a complex, ambitious and laborious task. Nonetheless, this KB infrastructure aims to present an approach to solve this problem. Figure 4 that depicts the system architecture of the CRC806-KB, which is an instance of an implementation of the here presented concept.

The presented KB has primarily a CRC 806 project internal scope, meaning only project participants can edit the KB. The system allows to store all sorts of data, information (metadata, structure) and knowledge (queries, filters, visualizations) about published and unpublished resources. Data, information and knowledge is gathered and created by the project participants, by editing the Wiki-based frontend in a collaborative, and thus sort of peer-reviewed, or at least peer-controlled or peer-aware approach. The resulting KB can then be queried through complex spatio-temporal queries, such as "show all archaeological sites, with artifacts classified as Aurignacien culture and located in northern Spain" for example. This query will yield a certain result set, that can be directly visualized on a web-based map, or shown in form of a table and even exported in many different formats, such as Excel, XML or JSON for example. On this basis, an infrastructure, that integrates available, already published, datasets and databases of interest to the research questions of the CRC 806, allows to enter and handle manually entered data from available publications into defined forms (schema based). It is also possible to build up a bibliographic data base of related relevant research publications, that can all be collaboratively edited, discovered and accessed through a single user friendly web application.

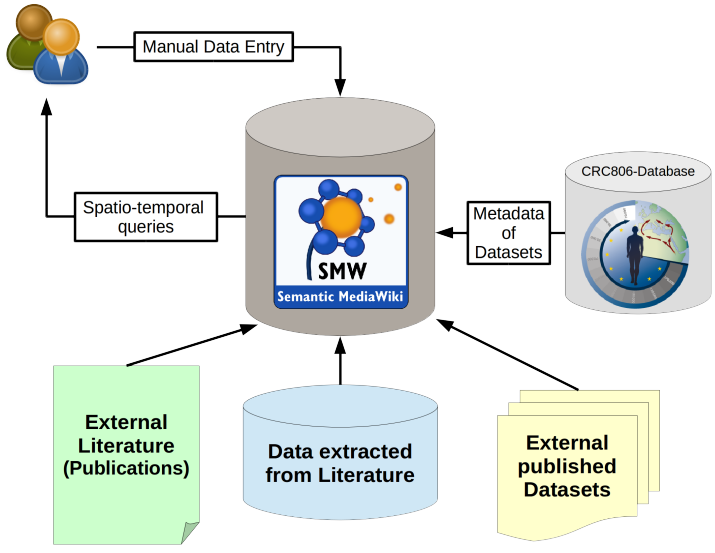

Figure 4. System architecture of the CRC806-KB. Source: [2].

3.1. Knowledge Management in SMW

In SMW information (structure or metadata) and knowledge (queries or algorithms) are managed based on semantic triples and properties, and queries upon those properties and triples. Information and data entry into the system is facilitated using Semantic Forms [47], sophisticated display of knowledge stored in the system is facilitated using Semantic Result Formats [48]. These techniques are briefly introduced in the following four sub-sections.

3.1.1. Semantic Triples and Properties

SMW's main feature is, that it enables MW to manage structured data. In SMW a datum (data item) is represented as a semantic triple. Semantic Triples are also the central concept of Semantic Web Technology (SWT) [49] and formalized as the Resource Description Format (RDF) [50]. A triple consist of a three-part structure: a subject, a predicate and an object [38]. An example would be:

```
Germany-Has capital-Berlin
```

where "`Germany`" is the subject, "`Has capital`" is the predicate (or relationship, or link), and "`Berlin`" is the object. In MW all content is stored in *wikitext* notation on wiki pages. This basic principle of MW also applies to SMW content. In SMW, the predicate is known as the "Property", and the subject is always the Wiki page on which the value is stored [38]. To encode the example triple in SMW would be to store the following string on the Wiki page of name "Germany":

```
[[Has capital::Berlin]]
```

This syntax, allows the SMW wikitext parser to capture the semantic triple in its data base, and make it available for queries. Subjects are pages, predicates are SMW properties and objects are variables or values, given as numbers or strings according to the defined format of the property. Properties in SMW can have different types, and it depends on the type if and how the above notated triple is displayed or rendered on the Wiki page. Further details on how to define properties, are given in the SMW documentation [23]. In summary, all SMW data and information content is stored via wikitext markup in the Wiki.

3.1.2. Queries

If structured data is stored, it is obviously desirable to be able to query this data. In SMW, queries on the structured data are facilitated from the ASK query language of SMW [23]. The syntax of this query language is similar to the syntax of annotations in SMW. This query language can be used on the SMW special page `Special:Ask`, in SMW concepts, and in inline queries[5].

SMW queries consists of two parts; (1) which pages (subjects) to select; and (2) What information (properties) to display about those pages. All queries have to state some conditions that describe what is asked for. You can select pages by name, namespace, category, and most importantly by property values. For example, the query:

```
{{#ask:[[Category:Countries]]|?  Has capital}}
```

Would yield a list of Countries and their Capitals stored in the Wiki. The first, "`[[Category:Countries]]`", is the filter—it defines which pages get queried; in this case, all pages in the category "`Countries`". The second part, after the "`|`", is called "printout", and selects the properties of the filtered pages (subjects) to display. In the example, all properties of "`Has capital`".

3.1.3. Semantic Forms

The Semantic Forms extension [47] provide a way to edit template calls within a Wiki page, where the templates are facilitated to store structured information in SMW. It thus complements SMW, by providing a structure for SMW's storage capabilities [38]. The concept of SF is based on the MW templating concept. MW templates can provide structure and the definition of the display of the structured content to Wiki pages. Thus, templates are useful for structuring the input of content to MW, and delivering a definition for the display of the content.

---

[5] https://semantic-mediawiki.org/wiki/Help:Semantic_search

3.1.4. Semantic Result Formats

The Semantic Result Formats (SRF) extension [48] provide additional result formats for SMW inline queries, to display query results (knowledge) in many formats, layouts and visualizations [38]. The version of SRF that is used in the here presented installation, includes 41 semantic result formats, that are available to visualize and export query results. These result formats cover almost any use case. There are result formats for calendars, timelines, charts, graphs and mathematical functions. On the extensions website [48], all result formats are listed and documented, the formats are organized in seven categories; misc, math, export, time, charts, tables, and graphs. See Figure 5 for a screenshot of the SMW Query interface including a list of available SRF to choose from.

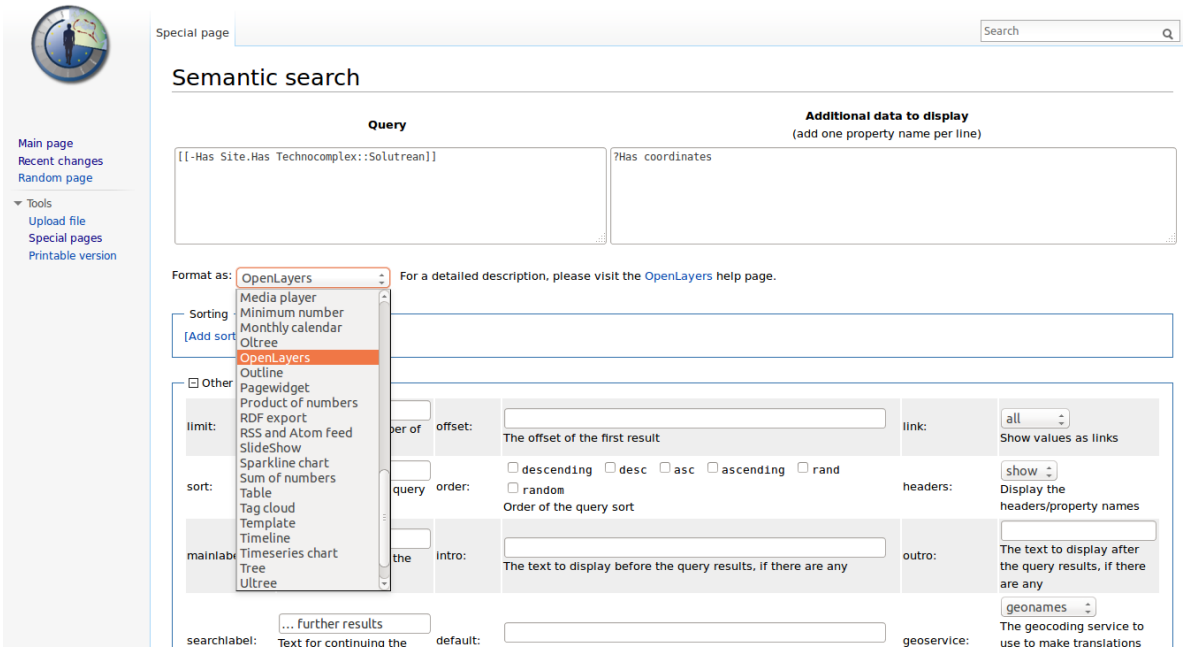

Figure 5. SMW Query interface with selection of different Result Formats.

3.1.5. Semantic Maps

A special SRF is the *Semantic Maps* extension [51], it allows to show query results, containing properties of special SMW type `Geographic Coordinates`. In Semantic Maps it is possible to use multiple mapping services. These include Google Maps (with Google Earth support), Yahoo! Maps, and OpenLayers [51]. In Figure 6, an example inline query, that produces a Semantic Map, showing all Sites with Technocomplex Solutrean, by its property *Coordinates*. The property `Coordinates` needs to be of type `Geographic Coordinates`, which is a special type defined by the SemanticMaps extension.

3.2. Data Entry

In SMW it is possible to define web forms for data entry. Those forms can consist of all standard HTML form fields, plus special input fields for SMWs own data types, for example a *map input* to define properties of the type *Geographic Coordinates* or a calendar input to define properties of the type *Date*.

A common use case for collecting data within the CRC 806 is to enter data from published literature. The data published in a traditional publication (e.g., Journal articles, Books, or Excavation reports) can be very heterogeneous. The idea is to provide data entry forms, annotating a publication resource with data. At first, the bibliographic metadata of the publication is entered into the Wiki including generation of a reference key for the publication, and used to link all information originating from this piece of literature.

	Age	Uncertainty	LabCode	SampleMaterial	Method
A Valina-GrN-20833-INQUA6747	31,730 31,600	2,800 250	GrN-20833 GrA-3014	bone from level V? charcoal	14C AMS
A du Blot (Cerzat)-Ly-564-Stage3 Arch-551	21,700 24,841	1,200 1,112	Ly-564	carbonaceous earth	C14
A du Blot (Cerzat)-Ly-565-Stage3 Arch-552	21,500 24,699	700 697	Ly-565	carbonaceous earth	C14
A du Mas Viel (St-Simon)-Gif-3281-Stage3 Arch-542	26,770 30,435	800 1,004	Gif-3281	bone	C14

Figure 6. Example map rendered by the Semantic maps Extension.

3.3. Data Integration

As mentioned in the introduction, the research of the CRC 806 is—at its core—of spatiotemporal nature. Time and space are the main integrating factors of the presented data base. The data is spatially integrated by its spatial extent. For GIS data the spatial extent is present intrinsically in the data format. For data, not given in a GIS data format, or not containing explicit geo coordinates, the spatial integration is facilitated by annotating spatial attributes with predefined regions or sites. Those translate into pre-defined bounding boxes, polygons (areas, regions) or point coordinates (sites). The same is implemented for temporal data, where the data can be annotated with predefined periods and events, which translate into time-spans (periods) between a start and an end date, or into simple dates (events).

4. Use Cases

The presented SMW based knowledge base was mainly build to collect and integrate data sources and datasets, as well as to produce geospatial datasets in GIS formats. Those data sources and datasets will be used as input for cartographic visualization, or in paleoenvironmental and archaeological modeling applications [9].

4.1. Contextual Areas

The Contextual Areas KB was developed for project partners of the B and C clusters of the CRC 806. The aim was to gather spatiotemporal archaeological information in one database, and to identify so called Contextual Areas in time and space.

For this KB a custom data integration workflow was developed and applied. Thus, a custom Python script for each of the datasets that generates DataTransfer XML, was implemented. The XML was then imported into the Wiki, using the DataTransfer extension. See Table 1 for an overview of the integrated archaeological datasets in this application.

The dataset and databases listed in Table 1, are all tables of dated remains or artefacts, that contain a date (point in time) including an error of the dating, and further information about the site (coordinates) where they were found, as well as the excavation context (location within the excavation trench, e.g., layer and section). A bibliography where the particular artefact with the according date was published, as well in some cases additional information on cultural (spatiotemporal) classifications of the artefact or remains. The custom data model of the Contextual Area KB consists of eight classes: *Artefact*, *Bibliography*, *Dataset*, *DatedAge*, *Layer*, *Region*, *Site*, and *TimePeriod*. Each of these eight classes describe certain objects with according defined properties. For example a *Site* has the Properties of

Name, *Latitude*, *Longitude*, *Altitude*, *Region*, and *Description*. This allows to ask spatiotemporal queries, like "give me all atrefacts of a TimePeriod from a Region". It is a new knowledge item, that was not available (that easy), to any project participant before.

Table 1. Integrated published archeological databases.

Database	# Records	# Sites	# Properties	Data Format	Source
INQUA DB	21,500	7238	54	MS Access	[52]
PACEA	6021	1209	26	CSV & MS Excel	[53]
Stage3 Arch	1896	380	20	MS Excel	[54]
Stage3 Faun	1912	502	24	MS Excel	[54]
CONTEXT	2874	441	31	MS Excel	[55]

Figure 7 shows screenshots of the *Contextual Areas* Wiki application. For example, in the upper left, a screenshot of a *TimePeriod* definition is given. In this case, it is the definition of the Aurignacien cultural period. On this *TimePeriod* knowledge item page, a map showing all *Site* objects containing *Artefacts* attributed with Aurignacien. Additionally, all *Artefacts* of this *TimePeriod* are listed in a broad table below the map.

The identification of Contextual Areas in the KB is simply facilitated by spatio-temporal queries. A simple contextual area is already shown on the Aurignacien map (see screenshot in Figure 7). These queries can be further refined spatially, by choosing smaller regions (smaller map extent), or temporally, by querying for smaller time intervals of 14C (and other methods) dates, as given in the KB.

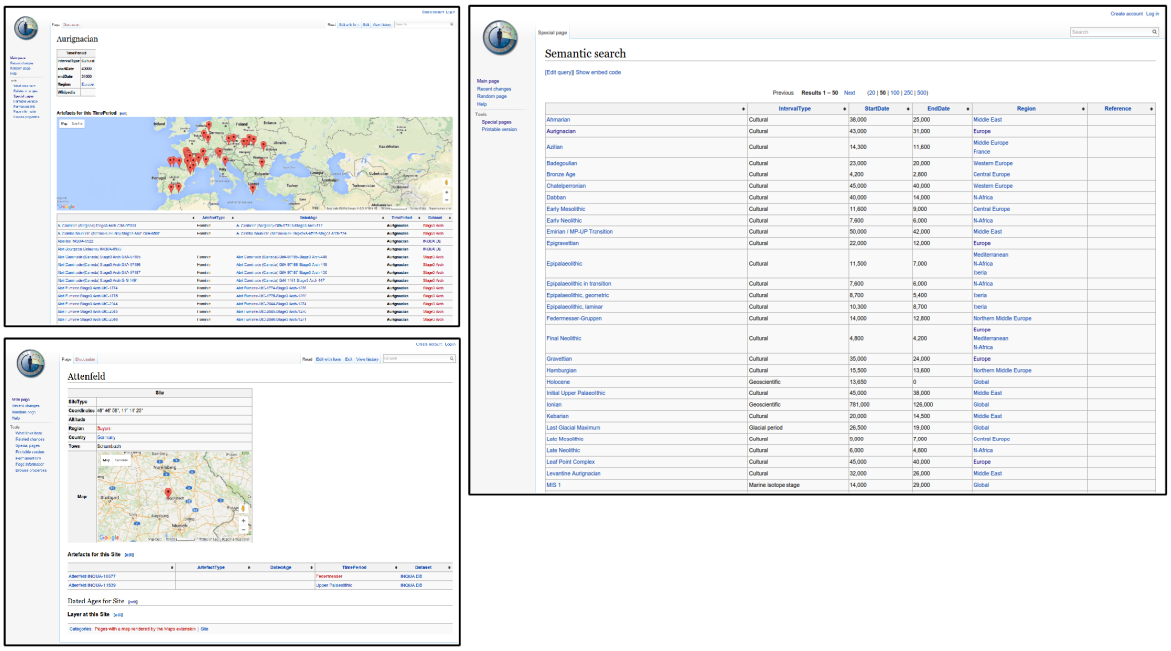

Figure 7. Screenshots of the Contextual Areas wiki interface.

4.2. Afriki

The Afriki KB was developed to assemble primary data of already published archaeological and palaeoecological results from Northeast Africa (Nile valley, Horn of Africa and African Rift valley) in the Late and Middle Pleistocene (0.012–0.78 Ma res. starting from MIS 19 to 3). The record of archaeological and geological proxy data is highly fragmented in this area [56].

In addition, available data are often not accessible in established repositories such as NOAA WDC Paleo or Pangaea [57]. Hence the compilation of data is essential to (re-)interpret data for

new approaches or different aspects [58]. In this context, it is also essential to take all restrictions and limits of previous studies in consideration to make the data comparable. Applied analytical methods and related age-depth models are also essential for the evaluation process of published scientific data. The amount of available data in so-called "grey" literature is enormous and has to be carefully evaluated on their robustness and partly (re-)processed to meet international standards. Often, the data have to be excerpt from figures or tables that are source of scientific interpretations (e.g., palaeoenvironment, palaeoclimate, evolution patterns, time models etc.). Furthermore, the names of the study sites are often transcribed from different languages or hold several synonyms for various reasons. Thus, we decided to use a semantic wiki to have the advantage of query-able structured data combined with the ability of web-based frontend for collaborative editing of the content [59].

Details of published and unpublished archaeological and geological sites/localities in East Africa are collected in the presented Wiki including their bibliographical reference. For example from sediment records, results from available sedimentological/chemical/biological proxy data (e.g., grain size, total organic carbon, stable isotopes, diatoms, ostracods, magnetic susceptibility) are copied into the database including their spatial resolution. Related dating samples (i.e., 14C, OSL, TSL) are also included with their metadata and lab-codes [59].

Technically, this Wiki instance was enhanced with a custom theme (i.e., layout), as shown in Figure 8. The theme was developed and customized by the project partners of the A3 project. The data collection in this KB instance, was also completely carried out by the A3 project. The assistance by the data management and GIS project Z2, was in providing the Wiki infrastructure and help in developing the data model and data queries including according visualizations.

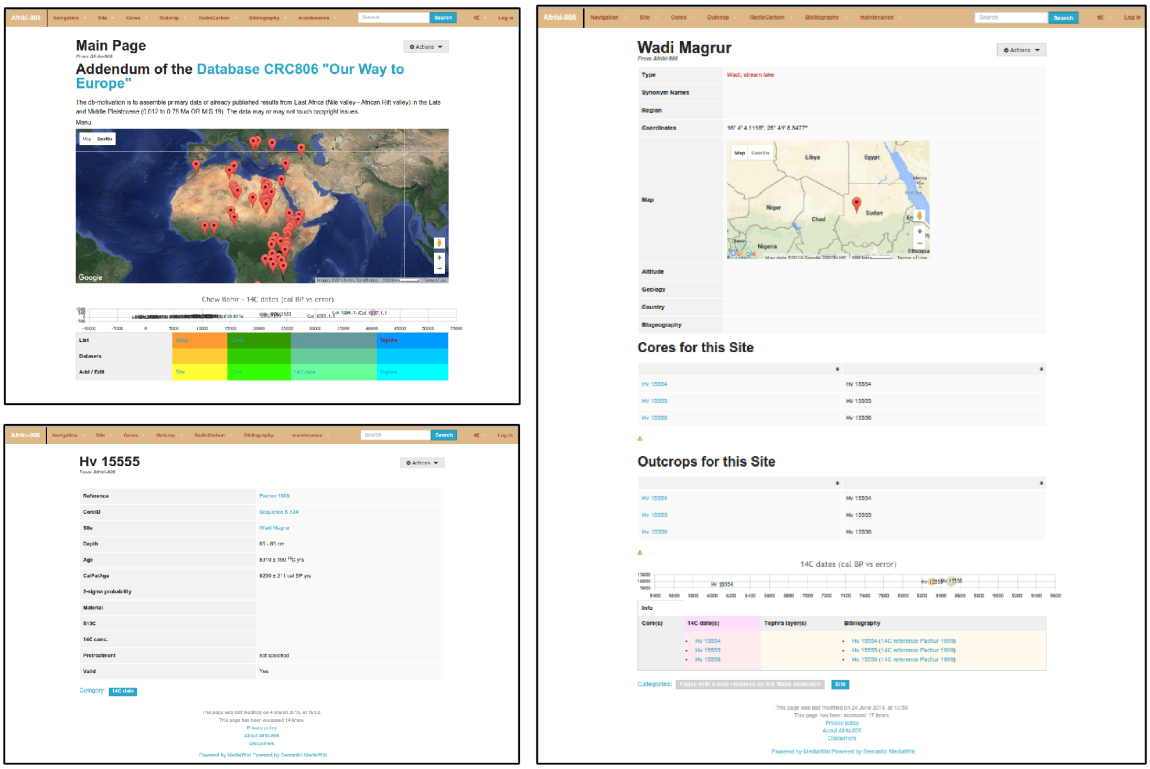

Figure 8. Screenshots of the Afriki-KB [59] interface.

The main information item handled in this application are analysis results of ostracods and other faunal and organic remains, which are dated and have an 14 C dating, optically stimulated luminescence (OSL) or electron spin resonance (ESR) dating (i.e., point in time), and originate from cores of lake sediments. These entail always a site (coordinates), as well as an identification code of the core (series). Additionally, the bibliographic information where the data was published is annotated

as an bibliography item. The data model of the application consists of 8 classes: *Bibliography*, *Site*, *14C Data*, *Core*, *FaunTaxon*, *OstracodTaxon*, *Outcrop*, and *TephraData*.

A feature that makes this KB instance interesting, is the application of temporal visualizations (see Figure 8), that allows to visualize dates on an interactive zoom-able timeline. Map visualizations of spatial annotated data was also implemented for this Wiki. These queries also generate new and easily accessible knowledge, that are updated if new data is integrated into the SMW based knowledge base.

5. Discussion and Conclusions

This study presents a concept and implementation of a web-based collaborative knowledge base system, based on semantic wiki technology. By presenting two use cases, it was shown that this technology is well suited to implement smaller project based web platforms that enable the project participants to collaboratively collect, edit, annotate, create and share data, information and knowledge.

The main problem of the evolving database is the reliability, validity and quality of the integrated data. All integrated datasets vary on each of these dimension. One of the most disputed information in those data sets is the central information on which these datasets are built. Those are the dated ages for artefacts, archaeological or sediment layers or the related age-depth models. Consequently, the provenance of the data is most important. Sufficient provenance information is needed to enable the researchers using the database, and to judge its data on an informed basis. To enable this, the original data source is provided for any dataset. A data source can be a scientific publication (bibliography) or a dataset. In case of a publication, the user is either referred to the bibliographic metadata of the publication including a PDF resource or to the publishers website containing the content of that publication, if existent. If the data source is of the type dataset, the dataset page has information about the original datasets, its source and according publications, as well as the schema mapping definition that mapped the data to the integrated data model.

The combination of a well equipped collaborative web platform facilitated by Mediawiki, the possibility to store and query structured data in this collaborative database, as well as the possibility for automated data import and data model development result in a powerful but flexible system to build a collaborative knowledge base.

A major downside of smaller project collaborative web applications, like the presented system, is its vulnerability to spam and hacking attacks. Several major spam attacks, as well as hacking attempts, forced us to ban access to the system from outside the university of Cologne's network (UKLAN). It was not possible for the author to handle the amount and severity of those attacks. The vulnerability to spam and hacking attacks is a major weakness of MediaWiki, we observed in many cases that these attacks were conducted by humans presumably working at click farms, and not only by automated spam bots. Thus it was nearly impossible to find a good balance between server hardening to prevent unwanted access, and usability of the application for the project participants.

The combination of a well equipped Wiki based collaborative web platform facilitated the possibility to store and query structured data in a collaborative database, as well as the possibility for automated data import and export. The data model development results in a powerful but flexible system that is able to build a collaborative knowledge base.

Author Contributions: C.W. developed the CRC806-KB system, designed and conducted the study, and wrote the manuscript. F.V. implemented the Afriki case study, wrote the according section about it, and helped with writing, editing and review of the manuscript. S.E.L. helped with the Afriki case study and helped with reviewing and editing the manuscript. G.B. secured funding for the data management project within the CRC 806, provided feedback on the infrastructure choices and facilitated to develop this idea.

Funding: The presented research was funded by the German Research Foundation (DFG) through the Collaborative Research Centre 806—"Our Way to Europe", Projects Z2 and A3, http://www.sfb806.de.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Richter, J.; Melles, M.; Schäbitz, F. Temporal and spatial corridors of Homo sapiens sapiens population dynamics during the Late Pleistocene and early Holocene. *Quat. Int.* **2012**, *274*, 1–4. [CrossRef]
2. Willmes, C. CRC806-Database: A Semantic E-Science Infrastructure for an Interdisciplinary Research Centre. Ph.D. Thesis, University of Cologne, Köln, Germany, 2016.
3. Willmes, C.; Kürner, D.; Bareth, G. Building Research Data Management Infrastructure using Open Source Software. *Trans. GIS* **2014**, *18*, 496–509. [CrossRef]
4. Willmes, C.; Yener, Y.; Gilgenberg, A.; Bareth, G. CRC806-Database: Integrating Typo3 with GeoNode and CKAN. In Proceedings of the 2nd Workshop on Datamanagement, University of Cologne, Cologne, Germany, 28–29 November 2014. [CrossRef]
5. Willmes, C.; Brocks, S.; Hoffmeister, D.; Hütt, C.; Kürner, D.; Volland, K.; Bareth, G. Facilitating integrated spatio-temporal visualization and analysis of heterogeneous archaeological and palaeoenvironmental research data. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **2012**, *I-2*, 223–228. [CrossRef]
6. Effertz, E. The funders perspective: Data management in coordinated programmes of the German Research Foundation (DFG). In Proceedings of the Data Management Workshop, University of Cologne, Cologne, Germany, 29–30 October 2010; pp. 35–38. [CrossRef]
7. DFG. *Sicherung Guter Wissenschaftlicher Praxis: Safeguarding Good Scientific Practice*; DFG: Bonn, Germany, 2006. [CrossRef]
8. DFG. *Recommendations for Secure Storage and Availability of Digital Primary Research Data*; Committee on Scientific Library Services and Information Systems—Subcommittee on Information Management, Deutsche Forschungsgemeinschaft, 53170 Bonn, Wissenschaftliche Literaturversorgungs—Und Informationssysteme (LIS); DFG: Bonn, Germany, 2009.
9. Willmes, C.; Becker, D.; Verheul, J.; Yener, Y.; Zickel, M.; Bolten, A.; Bubenzer, O.; Bareth, G. PaleoMaps: SDI for open paleoenvironmental GIS data. *IJSDIR* **2017**, *12*, 39–61. [CrossRef]
10. Fraser, M. Virtual Research Environments: Overview and Activity. *Ariadne* **2005**. Available online: http://www.ariadne.ac.uk/issue44/fraser/ (accessed on 23 October 2018).
11. Hey, T.; Trefethen, A.E. Cyberinfrastructure for e-Science. *Science* **2005**, *308*, 817–821. [CrossRef] [PubMed]
12. Ackoff, R.L. From Data to Wisdom. *J. Appl. Syst. Anal.* **1989**, *16*, 3–9.
13. Jennex, M. Re-Visiting the Knowledge Pyramid. In Proceedings of the 2nd Hawaii International Conference on System Sciences, Waikoloa, HI, USA, 5–8 January 2009; pp. 1–7. [CrossRef]
14. Rowley, J. The wisdom hierarchy: Representations of the DIKW hierarchy. *J. Inf. Sci.* **2007**, *33*, 163–180. [CrossRef]
15. Frické, M. The knowledge pyramid: A critique of the DIKW hierarchy. *J. Inf. Sci.* **2009**, *35*, 131–142. [CrossRef]
16. Willmes, C.; Bareth, G. A data integration concept for an interdisciplinary research database. In Proceedings of the Young Researchers forum on Geographic Information Science—GI Zeitgeist, Muenster, Germany, 16–17 March 2012; pp. 67–72.
17. Krötzsch, M.; Vrandečić, D.; Völkel, M. Semantic MediaWiki. In *The Semantic Web—ISWC 2006*; Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2006; Volume 4273, pp. 935–942. [CrossRef]
18. Alavi, M.; Leidner, D.E. Review: Knowledge management and Knowledge Management Systems: Conceptual Foundations and Research Issues. *MIS Q.* **2001**, *25*, 107–136. [CrossRef]
19. Baumeister, J.; Reutelshoefer, J.; Puppe, F. Know we: A semantic wiki for knowledge engineering. *Appl. Intell.* **2011**, *35*, 323–344. [CrossRef]
20. Fink, K.; Ploder, C. Balanced system for knowledge process management in SMEs. *J. Enterp. Inf. Manag.* **2009**, *22*, 36–50. [CrossRef]
21. Centobelli, P.; Cerchione, R.; Esposito, E. How to deal with knowledge management misalignment: A taxonomy based on a 3D fuzzy methodology. *J. Knowl. Manag.* **2018**, *22*, 538–566. [CrossRef]
22. Centobelli, P.; Cerchione, R.; Esposito, E. Aligning enterprise knowledge and knowledge management systems to improve efficiency and effectiveness performance: A three-dimensional Fuzzy-based decision support system. *Expert Syst. Appl.* **2018**, *91*, 107–126. [CrossRef]

23. SemanticMediawiki Contributors. Semantic MediaWiki—Free and Open-Source Extension to MediaWiki, 2018. Available online: https://semantic-mediawiki.org/ (accessed on 23 October 2018).
24. Vrandecic, D.; Krötzsch, M. Reusing Ontological Background Knowledge in Semantic Wikis. In Proceedings of the First Workshop on Semantic Wikis, Budva, Montenegro, 12 June 2006.
25. Bradtmöller, M.; Pastoors, A.; Slizewski, A.; Weniger, G.C. NESPOS—A digital archive and platform for Pleistocene archaeology. In Proceedings of the Data Management Workshop, University of Cologne, Cologne, Germany, 29–30 October 2010; pp. 13–18. [CrossRef]
26. Huvila, I. Being Formal and Flexible: Semantic Wiki as an Archaeological e-Science Infrastructure. In *Revive the Past. Computer Applications and Quantitative Methods in Archaeology (CAA)*; Zhou, M., Romanowska, I., Wu, Z., Xu, P., Verhagen, P., Eds.; Pallas Publications: Amsterdam, The Netherlands, 2012; pp. 186–197.
27. Huvila, I.; Huggett, J. Archaeological Practices, Knowledge Work and Digitalisation. *J. Comput. Appl. Archaeol.* **2018**, *1*, 88–100. [CrossRef]
28. Berners-Lee, T.; Hendler, J.; Lassila, O. The Semantic Web. *Sci. Am.* **2001**, *284*, 34–43. [CrossRef]
29. Uschold, M.; Gruninger, M. Ontologies and semantics for seamless connectivity. *ACM SIGMOD Rec.* **2004**, *33*, 58. [CrossRef]
30. Corcho, O.; Fernández-López, M.; Gómez-Pérez, A. Methodologies, tools and languages for building ontologies. Where is their meeting point? *Data Knowl. Eng.* **2003**, *46*, 41–64. [CrossRef]
31. Casellas, N. Methodologies, Tools and Languages for Ontology Design. In *Legal Ontology Engineering*; Number 1990; Springer: Berlin/Heidelberg, Germany, 2011. [CrossRef]
32. Chaudhary, D.; Yadav, P.K.; Singh, R.K.; Mitra, S.; Ghaziabad, I.P.E.C. Integrated Knowledge Base: An Approach to Knowledge Extraction. *Spec. Issue Int. J. Comput. Appl.* **2012**, *6*, 19–25.
33. Ziemba, P.; Jankowski, J.; Wątróbski, J.; Becker, J. Knowledge Management in Website Quality Evaluation Domain. In *Computational Collective Intelligence*; Núñez, M., Nguyen, N.T., Camacho, D., Trawiński, B., Eds.; Springer: Cham, Switzerland, 2015; pp. 75–85.
34. Guha, R.V.; Brickley, D.; Macbeth, S. Schema.Org: Evolution of Structured Data on the Web. *Commun. ACM* **2016**, *59*, 44–51. [CrossRef]
35. Vrandečić, D.; Krötzsch, M. Wikidata: A Free Collaborative Knowledgebase. *Commun. ACM* **2014**, *57*, 78–85. [CrossRef]
36. Schaffert, S.; Bry, F.; Baumeister, J.; Kiesel, M. Semantic Wikis. *Softw. IEEE* **2008**, *25*, 8–11. [CrossRef]
37. Wikimedia Fdn. MediaWiki, 2017. Available online: https://www.mediawiki.org/ (accessed on 23 October 2018).
38. Koren, Y. *Working with Mediawiki*; WikiWorks Press: San Bernardino, CA, USA, 2012.
39. Dengler, F.; Lamparter, S.; Hefke, M.; Abecker, A. Collaborative process development using semantic mediawiki. *Wissensmanagement* **2009**, *145*, 97–107.
40. Herzig, D.M.; Ell, B. *Semantic Mediawiki in Operation: Experiences with Building a Semantic Portal*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 114–128.
41. Jiang, G.; Solbrig, H.R.; Iberson-Hurst, D.; Kush, R.D.; Chute, C.G. A collaborative framework for representation and harmonization of clinical study data elements using semantic MediaWiki. *Summit Transl. Bioinform.* **2010**, *2010*, 11.
42. Boulos, M.N.K. Semantic Wikis: A comprehensible introduction with examples from the health sciences. *J. Emerg. Technol. Web Intell.* **2009**, *1*, 94–96. [CrossRef]
43. Hemler, S. Mobo Software, 2018. Available online: https://github.com/Fannon/mobo (accessed on 23 October 2018).
44. Heimler, S. Semantic MediaWiki Model Development through Object-oriented JSON Schema. In Proceedings of the SMW CON FALL 2014, Vienna, Austria, 1–3 October 2014.
45. Heimler, S. Schema–Driven Development of Semanitc MediaWikis. Master's Thesis, University of Applied Sciences Augsburg, Augsburg, Germany, 2015.
46. Galiegue, F.; Court, G. JSON Schema: Core Definitions and Terminology, 2013. Available online: http://json-schema.org/latest/json-schema-core.html (accessed on 23 October 2018).
47. Koren, Y.; Gambke, S. Semantic Forms MediaWiki Extension, 2015. Available online: https://www.mediawiki.org/wiki/Extension:Semantic_Forms (accessed on 23 October 2018).

48. Koren, Y.; Kong, J.H.; Gambke, S.; Dauw, J.D. Semantic Result Formats SemanticMediaWiki Extension, 2017. Available online: https://semantic-mediawiki.org/wiki/Semantic_Result_Formats (accessed on 23 October 2018).
49. Allemang, D.; Hendler, J. *Semantic Web for the Working Ontologist: Modeling in RDF, RDFS and OWL*, 2nd ed.; Morgan Kaufmann Publishers/Elsevier: Burlington, MA, USA, 2011.
50. Carroll, J.J.; Klyne, G. Resource Description Framework (RDF): Concepts and Abstract Syntax. 2004. Available online: https://www.w3.org/TR/rdf-concepts/ (accessed on 23 October 2018).
51. De Dauw, J. Semantic Maps MediaWiki Extension, 2017. Available online: https://github.com/SemanticMediaWiki/SemanticMaps/ (accessed on 23 October 2018).
52. Vermeersch, P.M. Radiocarbon Palaeolithic Europe Database v14, 2011. Available online: http://ees.kuleuven.be/geography/projects/14c-palaeolithic/index.html (accessed on 23 October 2018).
53. D'Errico, F.; Banks, W.E.; Vanhaeren, M.; Laroulandie, V.; Langlais, M. PACEA Geo-Referenced Radiocarbon Database. *PaleoAnthropology* **2011**, *2011*, 1–12.
54. Van Andel, T.; Davies, W. *Neanderthals and Modern Humans in the European Landscape During the Last Glaciation: Archaeological Results of the Stage 3 Project*; McDonald Institute Archaeological Research Monographs: Cambridge, UK, 2003; p. 265.
55. Böhner, U.; Schyle, D. Radiocarbon CONTEXT Database, 2006. Available online: http://context-database.uni-koeln.de (accessed on 9 April 2015).
56. Blome, M.W.; Cohen, A.S.; Tryon, C.A.; Brooks, A.S.; Russell, J. The environmental context for the origins of modern human diversity: A synthesis of regional variability in African climate 150,000–30,000 years ago. *J. Hum. Evol.* **2012**, *62*, 563–592. [CrossRef] [PubMed]
57. Diepenbroek, M.; Grobe, H.; Reinke, M.; Schindler, U.; Schlitzer, R.; Sieger, R.; Wefer, G. PANGAEA—An information system for environmental sciences. *Comput. Geosci.* **2002**, *28*, 1201–1210. [CrossRef]
58. Viehberg, F.A.; Just, J.; Dean, J.R.; Wagner, B.; Franz, S.O.; Klasen, N.; Kleinen, T.; Ludwig, P.; Asrat, A.; Lamb, H.F.; et al. Environmental change during MIS4 and MIS 3 opened corridors in the Horn of Africa for Homo sapiens expansion. *Quat. Sci. Rev.* **2018**. [CrossRef]
59. Viehberg, F.A.; Willmes, C.; Esteban, S.; Vogelsang, R. *A Semantic Wiki as Repository to Review Published Palaeo-Data in East Africa*; CRC806-Database; University of Cologne: Cologne, Germany, 2015. [CrossRef]