Integrating Agriculture-related Data Provided by Thematic Networks into a High Impact Knowledge Reservoir

Integrating Agriculture-Related Data Provided by Thematic Networks into a High Impact Knowledge Reservoir

Hercules Panoutsopoulos[1], Borja Espejo Garcia[2], Philip E. G. Verbist[3], Spyros Fountas[4], Pieter Spanoghe[5], Christopher Brewster[6]

Dpt. of Natural Resources Management and Agricultural Engineering, Agricultural University of Athens, Greece[1, 2, 4] {hpanoutsopoulos@aua.gr +30 2105294029[1], borjaeg@aua.gr +34 657975572[2], sfountas@aua.gr +30 2105294035[4]}
Belgium, Philip.verbist@gmail.com +32 479 43 09 44[3]
Dpt. of Plants and Crops, Faculty of Bioscience Engineering, University of Ghent, Belgium, Pieter.Spanoghe@UGent.be +32 9 264 60 09[5]
Institute of Data Science, Maastricht University, Netherlands, Christopher.Brewster@maastrichtuniversity.nl +31 (0)6 2705 4646[6]

Introduction

Knowledge is conveyed through digital artifacts that people create as part of their everyday activities. Making these artifacts widely available may involve the use of database systems able to handle their storage and management. These systems employ database models explicitly defining the internal database structure. Database models are specific types of data models, whose role is to formally describe entities, as well as their associations, and further enable data type specifications, relation definitions and constraints identification (Umanath and Scamell 2014). However, technology provides a wide spectrum of possibilities for digital artifacts creation and data generation. The abundance of available technologies and tools allow for production of data in a variety of formats. So, in many cases, knowledge is communicated through semi-structured and unstructured rather than structured data. This has direct implications with regard to data store systems adoption and database models. Relational database management systems (RDBMSs) are not the optimal solution for storage and management of semi-structured and unstructured data, especially when large volumes of these kinds of data need to handled. As a consequence, a new paradigm of database technologies and respective database models has emerged. (da Silva et al. 2015).

The aim of this book chapter is to illustrate initial attempts towards a formal definition of the concept of Thematic Networks, through a specially designed ontology, and a database model for storing Agriculture-related data into a centralized data store. More specifically, its focus is EURAKNOS, an EU-funded, H2020[1] research project collecting Agriculture-related data in various formats and making it accessible to any interested parties and stakeholders for further analysis or development. Data is made

[1] https://ec.europa.eu/programmes/horizon2020/en/what-horizon-2020

available as outputs of Thematic Networks, which, in the context, of the project, are conceptualized as Knowledge Reservoirs. The goal of EURAKNOS is to aggregate and further disseminate data with important information to end-users; in other words, build a so called “High Impact Knowledge Reservoir”. However, there is a wide range in types and formats of data that need to be considered. Therefore, efficient management of heterogeneity is a core challenge for EURAKNOS and the various Thematic Networks, in the agricultural sector, which cover a range of activities from dairy farms to agroforestry.

Given the above, this book chapter begins with an outline of the concept of Thematic Networks and the rationale behind them. After that, a description of EURAKNOS project takes place with emphasis on the methodology that has been employed for identifying, collecting and making available Thematic Network outputs of interest. The ontology of Thematic Networks, aimed to be used for the purpose of metadata definition, is presented next. From a technical perspective, implementation of the High Impact Knowledge Reservoir will be based on an appropriately-designed web-based platform. So, after an overview of the platform’s architecture, design-related considerations are discussed. Issues include: (i) retrieval and integration of outputs of Thematic Networks, (ii) the EURAKNOS data store and database model, (iii) search technologies to be considered, and (iv) search-related user interaction. The book chapter concludes with a summary of key points and steps to follow.

The Role of Thematic Networks as Multipliers of Agriculture-Related Knowledge

EURAKNOS is a Thematic Network. Thematic Networks are a particular format of Multi-Actor projects promoted by EIP-AGRI and funded by EU's Horizon 2020 programme, working on a specific theme. They bring people from both science and practice together to create useful, practical outputs. The concept of Thematic Networks is at the core of EURAKNOS. It has stemmed from the need to efficiently address Agriculture- and Forestry-related problems of ever-increasing complexity and identify novel solutions. The dynamic agricultural landscape calls for out-of-the-box thinking and challenges the top-down, linear model of knowledge transfer from research community to practitioners. Scientists can no longer be viewed as the only innovation brokers in the Agriculture value chain. This means that other stakeholders also have an important role to play. In other words, solutions to problems need to occur as an outcome of coordinated, joint efforts of various actors bringing different perspectives and values to the process and, consequently, contributing to synergies between research and practice.

The above described approach is core to what the agricultural European Innovation Partnership (EIP-AGRI) calls an "*interactive innovation model*"[2] and constitutes the foundation of Multi-Actor projects (EIP-AGRI 2017). The use of the term "*interactive*" is indicative of the way in which innovation is conceptualized as it is expected to occur from bottom-up processes involving seminal interactions among various stakeholders (e.g. researchers, farmers, advisors, SME and NGO representatives) and aims to further already existing knowledge. Thematic Networks are a particular format of Multi-Actor projects (EIP-AGRI 2016) and as such they:

- relate to specific themes (e.g. sustainable cropping systems, animal production systems, plant health, rural dynamics and policies, knowledge and innovation systems) and focus on real problems of practitioners;
- bring together partners with complementary backgrounds and expertise to collaborate, throughout the entire project lifecycle, and come up with innovative solutions;

[2] https://ec.europa.eu/eip/agriculture/en/eip-agri-concept

● collect, further develop, as well as disseminate ready-to-apply recommendations and practices and, thus, serve as multipliers of Agriculture-related knowledge.

The key characteristics and properties of Thematic Networks are summed up in the following definition, which has been developed with the aim to facilitate the ontology creation:

Thematic networks are multi-actor projects collecting existing knowledge[3] and best practices on a given theme, which relates to a domain. Their aim is to produce outputs, having a specific purpose and addressing a particular topic, which are of different kinds (e.g. text, software, images, audio, video, and datasets) and formats. Outputs target end-users (e.g. farmers, foresters, advisors) and are produced by output creators, engaged in the Thematic Network, which can be distinguished into organizations (e.g. universities, SMEs, NGOs, research institutes), having different kinds of involvement in the Thematic Network (participatory or leading), and individual actors (e.g. farmers, livestock breeders, advisors).

EURAKNOS: A meta-Thematic Network for Collecting, Compiling and Enabling Access to Existing Best Practices and Knowledge

The goal of EURAKNOS project[4] is to collect, store and further disseminate already created Agriculture-related data. It is a Thematic Network aiming to strengthen agricultural knowledge and promote innovation across Europe. By bringing together academic organizations, research institutes, advisory centers, government bodies, SMEs, NGOs and farmer organizations, EURAKNOS embraces the Multi-Actor approach and acts as an “umbrella” project aiming to connect all Thematic Networks. It can be considered as a Thematic Network about Thematic Networks, i.e. a meta - Thematic Network. Through the coordinated efforts of a consortium of 17 organizations from 11 EU countries, with extensive expertise in innovation in Agriculture and Forestry, it has the following objectives:

● Establishment of connections with existing Multi-Actor Thematic Networks and increase of their impact as creators and multipliers of Agriculture-related knowledge.

● Collection, analysis and evaluation, on the basis of well-established quantitative and qualitative criteria, of the knowledge and tools made available by Thematic Networks.

[3] In the context of EURAKNOS project, the term knowledge is used to denote the data that convey agricultural knowledge.
[4] https://www.euraknos.eu/

● Development of guidelines and technical specifications for a centralized repository of agricultural knowledge, as well as recommendations for future Thematic Network - related initiatives.

● Development of a platform consisting of the EURAKNOS Thematic Network data store, search engine, and a web-based environment for querying and accessing Agriculture-related data.

● Promotion, sustainability and longevity of the project by using fit-for-purpose communication and dissemination channels and also pursuing links with other Thematic Networks, research projects (at the EU and national levels), as well as educational/training programs.

Methodology for Agriculture-Related Data Collection, Storage and Dissemination

Methodology overview and involved steps

As mentioned earlier, a term core to the project is that of "Knowledge Reservoir" (KR). EURAKNOS considers all Thematic Networks as KRs. A KR is a collection of digital artifacts communicating practical, ready-to-apply knowledge and able to contribute to innovative solutions for sustainable Agriculture and Forestry. The aim of EURAKNOS is to build a High Impact Knowledge Reservoir (HIKR), namely a KR that uses a content structuring model tailored to targeted end-user needs. Activities for data collection, storage and dissemination are described in a four-phase methodology. The phases of EURAKNOS methodology are named: (i) Evaluating, (ii) Determining, (iii) Exploring, and (iv) Widening. Descriptions of the activities of each phase are provided in Table 1 below.

#	phase	involved activities
1	Evaluating	• Definition of qualitative and quantitative criteria, based on cultural, socio-economic and environmental aspects, for ranking Thematic Network outputs with regard to their applicability at the local, regional, national and EU level. • Evaluation of ongoing and completed Thematic Networks. Operational Groups[5] and other relevant Multi-Actor projects are taken into consideration. • Evaluation focused upon produced knowledge and tools, as well as communication and dissemination channels employed for reaching end-users and maximizing the impact of Thematic Networks. • Evaluation of employed (by existing Thematic Networks) Multi-Actor approaches with the aim to estimate the short, medium, and long-term involvement of different types of actors.
2	Determining	Development of recommendations and guidelines for: • collection, storage and sharing of ready-to-be-used agricultural knowledge; • setting up an HIKR, tailored to targeted end-users' needs, enabling easy access to high quality content; • communicating and disseminating innovative knowledge to end-users; • interoperable KRs in future Thematic Network efforts.

[5] According to EIP-AGRI (2014), an Operational Group "*consists of several partners with a common interest in a specific, practical innovation project and the people involved in the Operational Group should be from a diverse combination of practical and scientific backgrounds (e.g. farmers, scientists, representatives of the agri-business sector, etc.).*"

#	phase	involved activities
3	Exploring	• Development of ontology for describing the Thematic Networks domain. • Development of architecture and technical specifications for the EURAKNOS web-based platform. • Development of EURAKNOS web-based platform comprising: - a data repository, built on the basis of Open and FAIR data principles, for storing agricultural knowledge - related digital artifacts; - a search engine for executing queries on stored knowledge; - a user interface enabling querying and accessing available knowledge.
4	Widening	• Design and execution of communication, dissemination, exploitation and networking activities to maximize the impact of EURAKNOS. • Development of strategy for reaching a variety of audiences with special attention to EURAKNOS platform's end-users (e.g. farmers, foresters, advisors). • Synergies and connections with other Thematic Networks and Multi-Actor projects so as to: - contribute to the maximization of the impact of existing Thematic Networks; and - guide and structure future efforts in Knowledge Reservoirs' development. • Pursuit of sustainability and longevity of EURAKNOS by drawing links with training initiatives and educational programs at various sectors (e.g. academic, advisory, farm sectors) and levels (local, regional, national, and EU level). • Connection of EURAKNOS with EIP-AGRI to help develop links with existing and future Thematic Networks, Operational Groups and Focus Groups.

Table 1: Phases of EURAKNOS methodology, description of their scope and alignment with project WPs

After having mentioned the phases of the methodology that has been employed by EURAKNOS, as well as listed involved activities, it is helpful to proceed by presenting expected, from Thematic Networks' evaluation, outcomes and descriptions of what the recommendations for an HIKR should be about.

State-of-the-art review and recommendations for a High Impact Knowledge Reservoir

The first two methodology phases (namely, "Evaluating" and "Determining") set the ground for the development of technical specifications and construction of the EURAKNOS web-based platform. As mentioned above, the "Evaluating" phase involves a rigorous analysis of how Thematic Networks have addressed the issues of: (i) data storage and management, (ii) KR design, (iii) communication, dissemination and exploitation, and (iv) Multi-Actor engagement. To attain this goal, specially designed tasks (namely, "data impact assessment", "design of knowledge reservoirs", "dissemination, communication and information strategies" and "Multi-Actor approach"), have been executed. These tasks involved a desktop study and face-to-face interviews, of representatives of Thematic Networks, aiming to complement desktop studies' outcomes. Table 2 below lists all tasks involved in the specific methodology phase along with references to research tools and expected outcomes.

#	task name	task description	employed research tools and outcomes	
1	Data impact assessment	● Investigation and analysis of methods used for data generation and collection. ● Investigation of tools and technologies used for data storage. ● Investigation of data formats and types, as well as the content of available knowledge sources and tools. ● Estimation of produced knowledge completeness and identification of potential knowledge gaps.	desktop study	● Research of Thematic Network websites to get insights into content provision and information architecture. ● Evaluation of identified variations with regard to the different sectoral themes addressed by Thematic Networks (e.g. livestock, crop production, forestry).
			face-to-face interviews	● Interviews of Thematic Network representatives (e.g. coordinators, WP/Task leaders, technical infrastructure administrators).
2	Design of knowledge reservoirs	● Investigation and analysis of: - design of existing KRs; - database models used for structuring knowledge; - websites and data store systems of well-known international organizations (e.g. FAO, OECD, EFSA, etc.) contributing to provision and dissemination of agricultural knowledge. ● Analysis based on principles for open access to knowledge provided by initiatives such as the OpenAIRE project.	desktop study	● Research existing KRs for data and knowledge formats, and means for making it available to end-users. ● Research focused upon: (i) open source software solutions, (ii) format of available data and knowledge, (iii) search engines used, (iv) user interaction and experience.
			face-to-face interviews	● Interviews of Thematic Network representatives involved in the design of KRs (e.g. software/database engineers, technical infrastructure administrators).
3	Dissemination, communication and information strategies	● State-of-the-art review of communication and dissemination practices and tools. ● Special attention on how specific types of end-users (e.g. farmers, foresters and advisors), of particular interest to EURAKNOS, have been reached. ● Special consideration of social media.	desktop study	● Quantitative and qualitative evaluation of dissemination and communication tools and material by taking account of: particular needs of different end-user types, age, geographic location of end-users, frequency of use, etc.
			face-to-face interviews	● Interviews of Thematic Network representatives involved in the development and implementation of Thematic Networks communication, dissemination, and exploitation strategies.
4	Multi-Actor approach	● Focus on Multi-Actor approaches employed by existing Thematic Networks. ● Investigation of how Multi-Actor approaches have been conceptualized, put into motion and executed. Emphasis in the post-execution phase. ● Specific issues under consideration: - number and the type of actors (e.g. terms of represented sector and organization, geographic location); - methods of approaching different types of actors; - delivery of input by involved	desktop study	● Quantitative and qualitative analysis of information retrieved from Thematic Network websites.
			face-to-face interviews	● Interviews of Thematic Network partners and actors (in particular farmers, foresters and advisors) to further refine outcomes of the desktop study.

#	task name	task description	employed research tools and outcomes
		actors; - benefits in the short, medium and long term.	

Table 2: Tasks involved in the "Evaluating" phase of EURAKNOS methodology, employed research tools and expected outcomes

The "Evaluating" phase gives input to the "Determining" phase. This input is required for the needs of producing recommendations for building the EURAKNOS HIKR. In order to better illustrate the synergy between these two methodology phases, Table 3 below provides descriptions of the tasks of the "Evaluating" phase and links them to tasks of the "Determining" phase.

#	task name	task description	alignment with task of the "Evaluating" phase
1	HIKR data best practices and methodologies	● Definition of best methodologies and practices to generate, select, collect and store knowledge and data as outputs of Thematic Networks. ● Decisions about the data to be integrated in the EURAKNOS HIKR (decisions will be based on cultural, socio-economic and relevant environmental aspects). ● Recommendations for development of a database model and data structuring specifications for EURAKNOS HIKR.	Data impact assessment
2	Design of HIKR	● Recommendations and guidelines for the EURAKNOS HIKR design by taking account of: - targeted end-user needs; - technological issues (e.g. ingestion of data into the HIKR, development of data querying and access scenarios, data validation, user authentication and data security)	Design of knowledge reservoirs
3	HIKR dissemination, communication and exploitation strategies	● Identification and selection of tools to communicate project-related work, inform end-users and disseminate project outcomes. ● Exploitation of identified tools of best information, communication and dissemination channels to make the proposed innovation known to a wide range of potential end-users. ● Proposal of a new format for practice abstracts.	Dissemination, communication and information strategies
4	HIKR Multi-Actor approach	● Identification and selection of tools for Multi-Actor engagement and knowledge creation at different project phases, namely: conceptualization, initiation, execution and post execution of the project. ● Evaluation of stakeholder needs with regard to knowledge that needs to be produced and stored into the EURAKNOS HIKR.	Multi-Actor approach

Table 3: Tasks involved in the "Determining" phase of EURAKNOS methodology and alignment with tasks of the "Evaluating" phase

Towards an ontology of Thematic Networks

Knowledge Representation is dedicated to domain-specific knowledge modeling. Two widely adopted Knowledge Representation models are database models and ontologies. Although there are some differences (with many of them being just historical), they share a strong semantic heritage using different formalisms, such as logic, to build conceptual models of some subject matter. Ontologies are explicit specifications of a conceptualization (Gruber 1993) with their aim being to represent meaning rather than data. They are based on Description Logic, a low-level Knowledge Representation technique able to be documented through various markup languages such as RDF, RDFS, OWL and OWL-DL. It is important to remark that ontologies can provide solutions to data heterogeneity and interoperability problems and for that reason, they have played significant role in Semantic Web's development (Grimm et. al. 2011).

In order to develop the database model that will be used for data representation and storage in the EURAKNOS data store, an ontology of Thematic Networks has been decided to be created. Specifically, NOTICE (oNtology Of ThematIC nEtworks) is the ontology providing a formal description of the domain of Thematic Networks. Preliminary work on its development is presented in this book chapter. It helps towards identifying key entities related to the concept of Thematic Networks, their attributes and relations among these entities. Its creation, which has been based on the methodology proposed by Noy and McGuiness (2001), will facilitate the definition of metadata for the annotation of data in the EURAKNOS data store. Fig. 1 below illustrates the major steps involved in the ontology creation process.

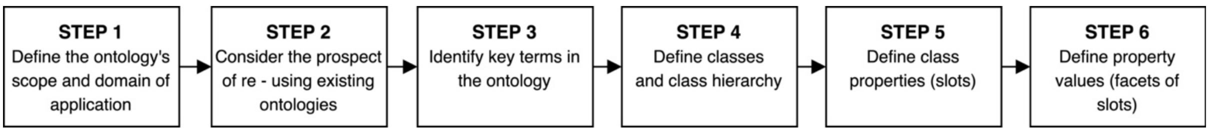

Fig. 1: Methodology employed for building NOTICE ontology (adapted from Noy & McGuiness, 2001)

To the best of our knowledge, no systematic efforts regarding formal, explicit descriptions of Thematic Networks (based on some kind of Knowledge Representation structure) exist. Therefore, the NOTICE ontology has been built from scratch. Identification of key terms, leading to the extraction of domain-related entities (which, in turn, are modeled as ontology classes and subclasses) and relations among them, has been based on the definition of Thematic Networks provided earlier. From that textual description, a number of statements containing references to domain-related entities, properties of those entities and relations among entities, have been derived. In these statements, entities and entity properties are denoted with bold characters and potential relations are indicated with italicized characters. More specifically:

- A **Thematic Network** *is about* a **theme**.
- The **theme** (of a Thematic Network) *relates to* a **domain**.
- An **output** of a Thematic network *targets* **end – users**.
- An **output** *serves* a **purpose**.
- An **output** *addresses* a **topic**.
- An **output** *has* a **format**.
- A **topic** *relates to* a **theme** (namely, the theme of the Thematic Network).
- An **output** *is produced by* an **output creator**.
- An **output creator** *is distinguished into* **individual actor** and **organization**.
- An **individual actor** *belongs to* an **organization**.
- An **organization** *has* a specific **type of involvement** (namely, leading or participating organization) in the Thematic Network.

The above statements have helped to come up with a set of ontology classes, subclasses and their properties all listed in Table 4 that follows.

class name	class definition	class properties	subclasses	subclass properties
Thematic Network	See definition provided in the section "The role of Thematic Networks as multipliers of Agriculture – related practical knowledge".	acronym fullName Url	-	-
Theme	The subject a Thematic Network is about.	title domain	-	-
Output Creator	The actor or group of actors whose activity/-ies have led to the creation of Thematic Network outputs.	name description country	**IndividualActor** A person who has been involved in the creation of a Thematic Network's output.	email type
			Organization An organization which has been involved in the creation of a Thematic Network's output.	areaServed contactPoint involvement type
Output	A digital artefact produced in the context of a Thematic Network.	abstract aggregateRating dateCreated	**AudioObject** An audio file.	duration format playerType transcript type

class name	class definition	class properties	subclasses	subclass properties
		inLanguage isAccessibleForFree keywords purpose size title	**Dataset** A body of structured information describing some topic(s) of interest.[6]	format issn measumentTechnique type variableMeasured
			DigitalDocument A document available in digital format with its content being mostly text.	format type
			ImageObject An image file.	height format type width
			PresentationDigitalDocument A file containing slides or used for a presentation.[7]	format type
			SoftwareApplication Software designed to perform a group of coordinated functions, tasks, or activities for the benefit of the user.[8]	availableOndevice downloadUrl format memoryRequirements operatingSystem softwareRequirements softwareVersion storageRequirements type
			VideoObject A video file.	duration format playerType transcript type videoFrameSize videoQuality
Topic	The subject an output of a Thematic Network is about.	title	-	-
EndUser	The human individual that uses any computing-enabled device/ appliance.[9]	type	-	-

Table 4: NOTICE ontology classes, subclasses and their properties

Definition of NOTICE classes, their subclasses and a number of their properties has been based on the following schema.org types: *Thing*, *CreativeWork*, *DigitalDocument*, *MediaObject*, *AudioObject*, *ImageObject*, *VideoObject*, *PresentationDigitalDocument*, *SoftwareApplication*,

[6] Definition provided by schema.org: https://schema.org/Dataset
[7] Definition provided by schema.org: https://schema.org/PresentationDigitalDocument
[8] Definition provided by Wikipedia (https://en.wikipedia.org/wiki/Application_software)
[9] Definition provided by Technopedia: https://www.techopedia.com/definition/610/end-user

Organization and *Person*. As mentioned in its official website[10], schema.org offers a vocabulary, able to be used with various encodings, for the definition of entities, relationships between entities and actions. Definition of the different kinds of digital artifacts, produced as Thematic Network outputs, has been based on the list of media types (termed as Multipurpose Internet Mail Extensions or MIME types) provided by the Internet Assigned Numbers Authority[11]. Types included in this list are: *application*, *audio*, *example*, *font*, *image*, *message*, *model*, *text* and *video*. In order to propose an as inclusive as possible list of Thematic Network output kinds, specifically tailored to the NOTICE ontology's design needs, the {*application*, *audio*, *image*, *text*, *video*} sublist of MIME types has been used and enriched with the *presentation* and *dataset* types. Table 5 below lists all kinds of Thematic Network outputs distinguished in NOTICE. These different output kinds are modeled as subclasses of the "Output" class. It needs to be stressed that each output serves a specific purpose (namely, *access to data*, *best practice presentation*, *communication*, *decision support*, *dissemination*, *educational material*, *innovative practice presentation*, *training material*), which consequently leads to a further division of each kind of outputs to different purpose-related types. In other words, each subclass of the "Output" class has a property named "type" assigned to it with its values depending on the purpose that the output serves.

subclasses of the "Output" class	values of "type" property
AudioObject	{advertising podcast, educational/training podcast, event capturing podcast, informational podcast, interview, on-demand seminar, tutorial}
Dataset	{auditory data, crop-related data, geospatial data, graph-related data, imagery data, input-related data, network-related data, temporal data, textual data, video data, yield-related data}
DigitalDocument	{article in conference proceedings, best practice guide, book, booklet, chapter in edited volume, deliverable report, factsheet, handbook/manual, journal article, milestone report, newsletter, practice abstract, press release, spreadsheet, review document, technical article, tutorial}
ImageObject	{chart/graph, figure/image, infographic}
PresentationDigitalObject	{demonstration, educational/training presentation, informational presentation, tutorial}
SoftwareApplication	{AI software, business software, data repository/database, decision support

[10] https://schema.org/
[11] https://www.iana.org/

subclasses of the "Output" class	values of "type" property
	tool, educational/training software, Farm Management Information System (FMIS), game, scientific software, simulation}
VideoObject	{advertising video, demonstration video, educational/training video, event capturing video, informational video, interview, testimonial, tutorial}

Table 5: Subclasses of the "Output" class and respective values of the purpose-related "type" property

Distinguishing different types of end-users is also important given that specific platform functionalities and solutions will be able to be provided given each user type's needs. On this basis, the following end-user types are defined: *adviser*, *bee keeper*, *consumer*, *distributor*, *entrepreneur*, *farmer*, *fisherman*, *forester*, *instructor*, *livestock breeder*, *policy maker*, *researcher*, *student*, *trader*, *trainer*, and *university professor*. As far as relations between classes are concerned, these are summarized in Table 6 that follows.

referencing class	relation	referenced class
ThematicNetwork	*engages*	OutputCreator
	isAbout	Theme
	isCoordinatedBy	Organization
Topic	*relatesTo*	Theme
Output	*addresses*	Topic
	isProducedBy	OutputCreator
	targets	EndUser
IndividualActor	*affiliatedWith*	Organization

Table 6: NOTICE ontology relations and involved classes

All the above are summarized in the graphic representation of NOTICE ontology illustrated in Fig. 2 below. Yet, as mentioned earlier, presented work on NOTICE ontology is preliminary. There are issues that need to be further taken care of. More specifically, there needs to be establishment of links between the NOTICE ontology and widely adopted, Agriculture-related ontologies and vocabularies for the needs of providing values to the domain and theme of Thematic Networks, as well as the topic of Thematic Network outputs.

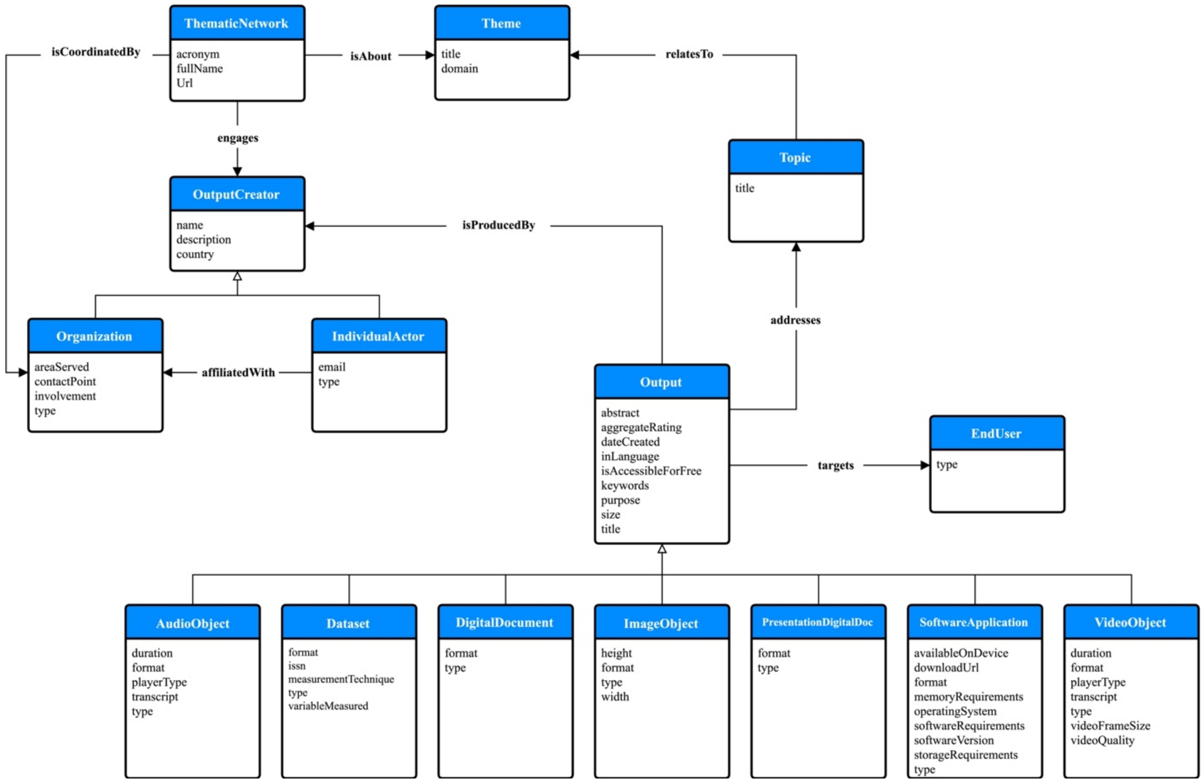

Fig. 2: The NOTICE ontology

Architecture of a web-based platform for Agriculture-related data

Architecture overview

The web-based platform constitutes the technical implementation of EURAKNOS HIKR. Platform description is based on a three-layer architecture consisting of, from bottom to top of the stack, the data persistence layer, the application layer and the presentation layer. The data persistence layer is responsible for data storage and management. It is the layer where the data store is. The application layer is concerned with the actual logic of the platform. This is the layer where the EURAKNOS search engine is. It handles data exchange between the presentation and data persistence layers and does not have to do with data persistence or how data is actually displayed to the end-user. However, it may well support multiple user interfaces (for instance, a web application and a mobile app). Finally, the presentation layer (in other words, the user interface) is what end-users see. It is the part of the application that end-users interact with. It displays data and receives requests from end-users. This layer is not concerned with any logic and it just relies on data. In order to make this stack of layers work, every layer needs to provide an Application Programming Interface (API) so that the other layers can communicate with it. So, for example, the EURAKNOS Search Engine may provide a REST Service and an OpenAPI (Swagger) specification on which a user interface may be built. An overview of EURAKNOS platform's architecture is provided in Fig. 3 below.

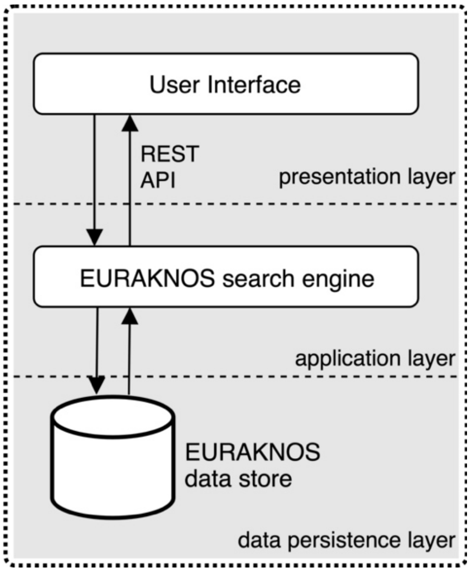

Fig. 3: Architecture of EURAKNOS web-based platform

Issues related to data collection and ingestion into the EURAKNOS data store

EURAKNOS aims to collect and organize agricultural knowledge - related outputs that are made available by existing Thematic Networks. These outputs have so far been developed and managed in a decentralized and unstructured way. So, in order to make data in the EURAKNOS data store searchable and identifiable, indexes need to be built. Apart from that, collection of large amounts of unstructured data will need to be based on some kind of automated process. Thematic Network outputs need to be aggregated into a single structured search index so that it can be searched. This is where web crawlers come into play. A web crawler is a program or automated script that browses the World Wide Web. It is usually employed to collect information, but it can also be used for automating maintenance tasks, such as checking for dead links on a website or testing websites for errors.

A point to consider is that it is important not to "harm" the website that is being crawled. To this end, politeness policies are used to prevent the crawled website from being heavily affected. So, a polite web crawler should:

- respect the rules defined in the website's robots.txt file;
- avoid degrading the performance of the website it is crawling;
- identify itself and its creator with contact information.

Humans are able to easily understand the content of a web page by just seeing it. However, it is not that easy for search engines to also interpret that same raw data. In order to help this interpretation process, a crawler collects metadata from the website. Search indexes can use metadata to interpret and link documents. The most obvious metadata provided by a web page are title-tags, the meta description and keywords. It needs to be stressed that, quite often, web pages also provide descriptions for images. In the case of lack of metadata, web crawlers can resort to other ways of data interpretation.

EURAKNOS data store and database model

Thematic Networks compile information in many different ways. Therefore, it is hard to define a strict schema for data collection. One way to address this problem with an RDBMS is to employ the EAV/CR model (namely, Entity Attribute Value with Classes and Relationships). RDBMSs are highly consistent, but this comes at the cost of not being able to horizontally scale. NoSQL systems, on the other hand, allow for horizontal scalability, yet, at the cost of consistency. Horizontal scaling refers to use of clusters of commodity hardware to store data, with each piece of hardware being responsible for the execution of processes, such as look-ups and read/write operations, on data stored in it (Lake and Crowther 2013). This specific capacity of NoSQL data stores is pointed out in the definition of Cattell (2010) according to which NoSQL systems are "*designed to provide good horizontal scalability for simple read/write database operations distributed over many servers*" (p. 12). However, advantages of NoSQL systems go beyond scalability-related issues. According to Vaish (2013), NoSQL data stores allow for schemaless data representation, which means that application developers are able to dynamically integrate changes into their design without needing to define a fixed data structure in advance. Schemaless data representation reduces development time given that data access can be handled by application code rather than complex SQL queries. As a consequence, there is potential to develop applications able to efficiently respond to various workloads and deliver results very quickly.

The solution intended to be adopted for the deployment of the EURAKNOS data store is MongoDB. MongoDB[12] belongs to the document-oriented category of NoSQL systems. By being a specific type of NoSQL data stores, document-oriented databases allow for adoption of a dynamic/changeable schema, or no schema at all, a characteristic that makes them ideal for storage of content changing over time (Vaish 2013). MongoDB is an open-source database management system providing high read and write throughput and the ability for horizontal scaling and automatic failure recovery (Banker 2012). MongoDB has become popular mostly because of its capacity to efficiently represent and retrieve hierarchically structured information without need for execution of resource-intensive table joins (Banker 2012). Moreover, according to Banker (2012), MongoDB supports ad hoc queries[13] and data indexing. It also provides automatic data replication, which means distribution of data across nodes of a cluster with the aim to eradicate data loss due to hardware or network failure. Distribution of data across nodes is internally handled by a mechanism called auto-sharding.

The model employed by MongoDB for data storage is called "document" and is based on the JSON (i.e. JavaScript Object Notation) data exchange format. More specifically, for performance optimization reasons, MongoDB makes use of a binary representation of JSON called BSON (Lake and Crowther 2013). BSON documents can have a maximum size of 16 MBs. However, larger documents can also be stored with the help of the GridFS API. MongoDB's document can be better conceptualized if considered as the counterpart of a table row in a RDBMS. A document is made up of property names and values. Values can be of any BSON compatible data type (e.g. string, number, Boolean, date), as well as array, document (called embedded document) and array of documents. Thus, MongoDB can

---

[12] https://www.mongodb.com/

[13] According to Technopedia (https://www.techopedia.com/definition/30581/ad-hoc-query-sql-programming), an ad hoc query is "*a loosely typed command/query whose value depends upon some variable. Each time the command is executed, the result is different, depending on the value of the variable.*"

offer great flexibility for modeling complex data structures (Banker 2012). A "collection" is a group of documents and can be considered the equivalent of a relational database's table. MongoDB allows for documents with different fields in the same collection.

Based on NOTICE ontology, the database model that is proposed for the Agriculture-related data representation and storage is shown in Fig. 4 below.

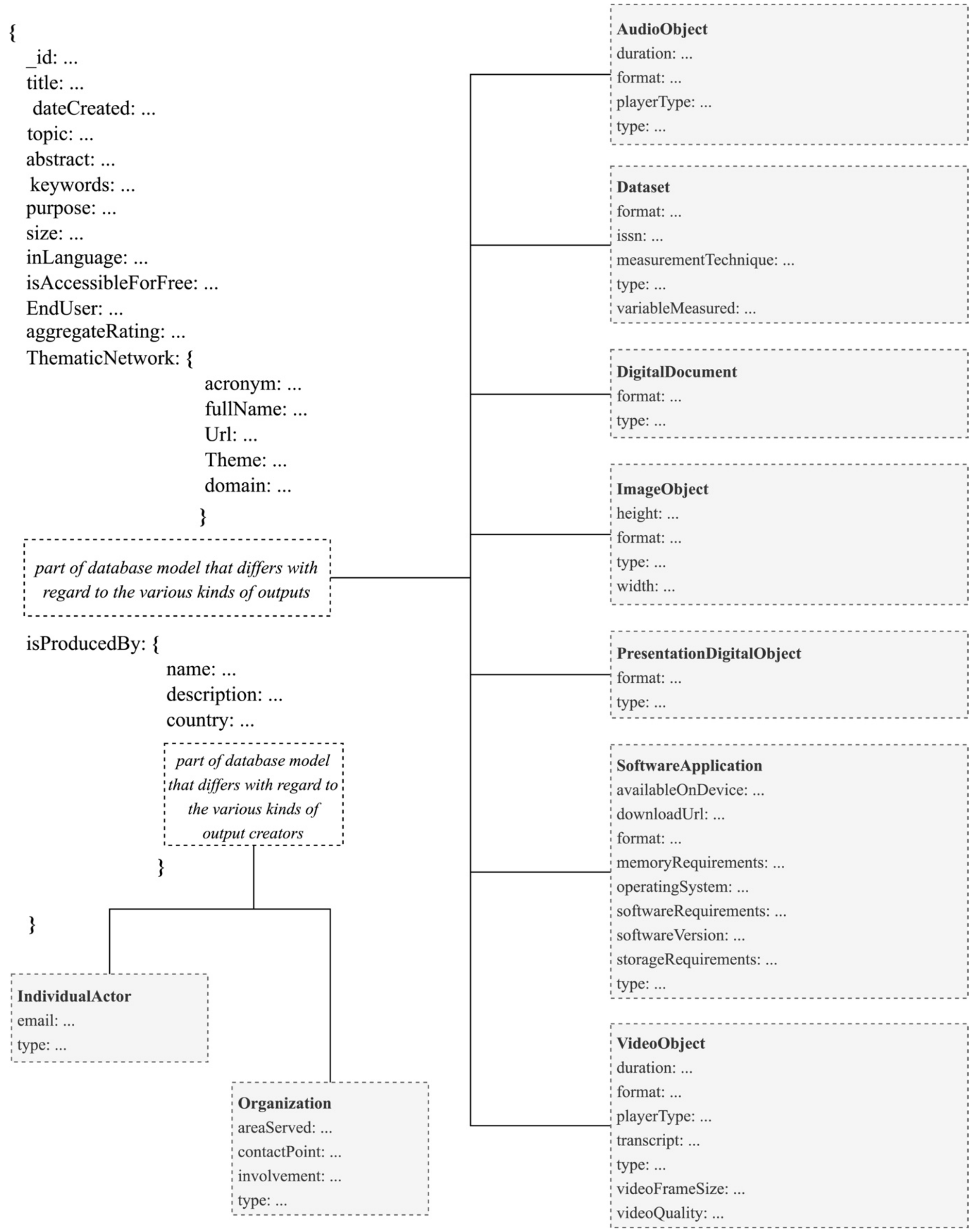

Fig. 4: EURAKNOS database model

Searching for Agriculture – related practical knowledge

The primary aim of a search engine is to retrieve the most relevant documents to user queries, excluding other general content (Brin and Page 1998). To achieve this goal, a search engine usually instantiates crawlers that visit links appearing in web pages and download documents represented by those links. Afterwards, an indexer parses the content of the documents (either textual or binary) and organizes it with the help of an index. Indexing is the method of storing data in index files, namely in a format helping fast and efficient text searching. There are different index types each of which has strengths and weaknesses. The most popular is the inverted index, which is a data structure that stores mappings from terms to a set of documents. Apache Lucene[14] is the most popular open source library for document indexing. By also enabling query submissions it allows to retrieve query-matching results. The index can be stored in the file system or memory and can be used to search the web, databases and ontologies. Prior to storing data in index files, data is analyzed and parsed. Given Lucene's efficiency and precise search algorithms, it is widely adopted for being at the core of a search engine. The effectiveness of a search engine can be measured by two popular performance metrics, namely precision (i.e. fraction of retrieved documents that are relevant) and recall (i.e. fraction of relevant documents that have been retrieved).

There is no right or wrong choice of search engine technology. It depends on application requirements and based on them, the most efficient solution can be determined. Among available open source solutions, the currently leading one, also qualifying for deployment of the EURAKNOS search engine, is Elasticsearch[15]. Elasticsearch was first released in 2010 and is built upon the Apache Lucene library. It is stable and has a well-documented reference guide. Its core functionality has been enriched beyond simple text indexing and searching. Over the years, features such as faceting (currently known as aggregations), field collapsing and language detection have been included. Elasticsearch exposes a REST API that allows, among others, to query, delete documents, as well as create and manage indexes by using the HTTP GET, DELETE, POST and PUT methods. As far as the output format of query results is concerned, Elasticsearch makes use of the JSON format. As far as scalability is concerned,

---

[14] https://lucene.apache.org/

[15] https://www.elastic.co/products/elasticsearch

Elasticsearch makes use of a modern cluster configuration instead of the old-fashioned Master-Slave architecture.

Interacting with EURAKNOS web-based platform

By the time this book chapter was authored, some preparatory work on the platform's back-end design had been done without, however, any final decisions on user interaction. So, the aim of this paragraph is to highlight interaction-related issues. Designing interaction with the platform is heavily focused on decisions about facilitating input from the end-user and delivery of results. On one hand, the data store will integrate digital artifacts of various formats and types waiting to be searched by different end-user types. On the other hand, end-users come to interact with the system with the expectation to get what they are looking for by bringing along their own presumptions and experiences from interactions with other applications. These critical issues need to be carefully considered for the needs of designing easy-to-use input and output interfaces providing straightforward prompts and the right amount of clearly structured information.

As far as user input is concerned, the most commonly employed design pattern is that of a search box accompanied by a dedicated search button. Enabling users to submit free-text queries is the most commonly employed practice providing great degree of freedom. In this case, use of tools able to facilitate query syntax and submission needs to be investigated. According to Rosenfeld et al. (2015), such tools are spell checkers which automatically correct misspelled query terms; stemming tools for retrieving results that contain terms sharing the same stem with terms in queries; Natural Language Processing tools having the capacity to provide results after syntactic analysis of submitted queries; controlled vocabularies and thesauri enabling search for and retrieval of results containing semantically similar terms; and autocompletion/autosuggestion tools able to significantly enhance the query development process through offer of query completion/matching options.

Despite the intuitive nature of free-text search and the advanced features that sophisticated add-ons can provide, yielded results may not always be those expected for. To this end, it is very important to investigate ways of further supporting end users' content searching endeavors. A solution towards this direction is to provide advanced search functionalities leading to more specific queries, and, thus, limiting the scope of search, through exploitation of logical operators. However, use of this feature should be opted by the user (e.g. in cases where difficulties are faced with regard to retrieving the required results). So, in terms of design, access to this kind of functionality should be made available through separate interfaces. In this context, an option worth investigating is that of interface segmentations called search zones (Rosenfeld et al. 2015), which offer content search functionalities, such as filters, targeting at different user type needs.

Design of search results display is also critical especially if diversity of Thematic Network outputs is considered. A question demanding attention within this context relates to the kind and amount of results-related information that needs to be displayed to the end-user. As Rosenfeld et al. (2015) point out, it is necessary to strike a balance between display of representational information (for instance, title, creator and date of creation of the obtained result), which better addresses the needs of users who are clear about what they are looking for, and descriptive information (in other words, information like the one mentioned above plus content summary and maybe some keywords) aiming to help users not quite certain of the results they need. In addition to all the above, it should not be neglected that there also needs to be a focus on the order of results display. Sorting and ranking are two popular options with the former appearing to be appropriate in cases where the user needs to make decisions or take actions and the latter serving better understanding establishment and learning purposes (Rosenfeld et al. 2015).

Discussion and Further Steps

The aim of this book chapter has been to present an ongoing systematic effort for collecting and making available Agriculture-related data. This process takes place in the context of EURAKNOS, a project that specifically focuses upon data provided by Thematic Networks. EURAKNOS is a Thematic Network itself aspiring to provide a single-point-of-access for Agriculture-related data retrieved by other, existing Thematic Networks. For this purpose, a specially designed methodology is under implementation. However, there are many challenges that need to be solved along the way to a unified collection of Thematic Network outputs. For instance, there is high diversity in the themes that have been addressed by Thematic Networks, as well as in types and formats of produced outputs. It should also be stressed that no single technology has been used for the development of Knowledge Reservoirs. More than that, no consensus exists with regard to employed database modeling approaches. As a consequence, there is a very fragmented view of the way in which knowledge has been represented and structured in the context of existing Thematic Networks.

Collecting agricultural data and providing access to it through a centralized data store requires a robust workflow and flexible database model. Stiff relational database models cannot be considered as adequate for representing the highly diversified outputs of Thematic Networks. Apart from that, EURAKNOS also aims to frame a context for future Thematic Networks by proposing, among others, an appropriate database model and system architecture along with fit-for-purpose technology recommendations. Steps towards this direction will be feasible to be made on the basis of NOTICE ontology and the EURAKNOS database model. Given that, work that needs to be further performed should be concerned with:

- Establishment of links between NOTICE ontology and well-known Agriculture-related ontologies and vocabularies for the needs of providing values to the domain and theme of Thematic Networks, as well as the topic of outputs from Thematic Networks.
- Facilitation of both human- and machine-enabled access to data through compliance with the FAIR (i.e. "Findable", "Accessible", "Interoperable", and "Reusable") guiding principles (Wilkinson et al. 2016) and open data standards - related initiatives such as the OpenAIRE project[16].
- Development of technical specifications based on a rigorous theoretical basis so as to assemble a robust system and take full advantage of cutting-edge software solutions like MongoDB and Elasticsearch.
- Finalization of interaction-related issues and architecture of provided information, given the various targeted end-user types and their respective needs.

References

Agricultural European Innovation Partnership 2014. EIP-AGRI Operational groups: Turning your idea into innovation. Retrieved from https://ec.europa.eu/eip/agriculture/sites/agri-eip/files/fact-sheet- operational-groups_en.pdf

Agricultural European Innovation Partnership 2016. Thematic Networks under Horizon 2020: Compiling knowledge ready for practice. Retrieved from https://ec.europa.eu/eip/agriculture/sites/agri-eip/files/ eip-agri_brochure_thematic_networks_2016_en_web.pdf

Agricultural European Innovation Partnership 2017. Horizon 2020 multi-actor projects. Retrieved from https://ec.europa.eu/eip/agriculture/sites/agri-eip/files/eip-agri_brochure_multi-actor_projects_2017_en_web.pdf

---

[16] https://www.openaire.eu/

Banker, K. 2012. MongoDB in action. Manning Publications Co., New York. USA

Brin, S. and Page, L. 1998. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In Seventh International World-Wide Web Conference (WWW 1998), Brisbane. Australia

Cattell, R. 2010. Scalable SQL and NoSQL data stores. Acm Sigmod Record 39 (4): 12-27.

da Silva, A. R., Silva, A., and Ribeiro, A. 2015. Data modeling and data analytics: a survey from a big data perspective. Journal of Software Engineering and Applications 8 (12): 720-726.

Grimm S., Abecker A., Völker J., and Studer R. Ontologies and the Semantic Web in Domingue J., Fensel D., and Hendler J.A. [eds.] 2011. Handbook of Semantic Web Technologies. Springer, Berlin, Germany.

Gruber, T.R. 1993. Toward principles for the design of ontologies used for knowledge sharing. Technical Report KSL 93-04, Knowledge Systems Laboratory, Stanford University.

Lake, P., and Crowther, P. 2013. Concise guide to databases. Springer, London. UK

Rosenfeld, L., Morville, P., and Arango, J. 2015. Information architecture: for the web and beyond. O'Reilly Media, Inc., Sebastopol, CA.

Noy, N. F., and McGuinness, D. L. 2001. Ontology development 101: A guide to creating your first ontology. Retrieved from https://protege.stanford.edu/publications/ontology_development/ontology101.pdf

Umanath, N. S., and Scamell, R. W. 2014. Data modeling and database design. Delmar Learning, New York, USA.

Vaish, G. 2013. Getting started with NoSQL. Packt Publishing Ltd, Birmingham. UK