Environmental Modelling & Software 84 (2016) 494–504

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Analysis of Big Data technologies for use in agro-environmental science

Rob Lokers*, Rob Knapen, Sander Janssen, Yke van Randen, Jacques Jansen

Alterra Wageningen UR, P.O. Box 47, 6700 AA Wageningen, The Netherlands

ARTICLE INFO

Article history:
Received 21 January 2016
Received in revised form
27 July 2016
Accepted 29 July 2016
Available online 25 August 2016

Keywords:
Big Data
Semantics
Data integration
Interdisciplinary research
Agriculture
Forestry

ABSTRACT

Recent developments like the movements of open access and open data and the unprecedented growth of data, which has come forward as Big Data, have shifted focus to methods to effectively handle such data for use in agro-environmental research. Big Data technologies, together with the increased use of cloud based and high performance computing, create new opportunities for data intensive science in the multi-disciplinary agro-environmental domain. A theoretical framework is presented to structure and analyse data-intensive cases and is applied to three case studies, together covering a broad range of technologies and aspects related to Big Data usage. The case studies indicate that most persistent issues in the area of data-intensive research evolve around capturing the huge heterogeneity of interdisciplinary data and around creating trust between data providers and data users. It is therefore recommended that efforts from the agro-environmental domain concentrate on the issues of variety and veracity.

1. Introduction

Societal challenges (e.g. food security, ecosystem restoration, climate change, resource use efficiency as captured in the Sustainable Development Goals (https://sustainabledevelopment.un.org/topics/sustainabledevelopmentgoals) and EU's societal challenges (https://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges) require more and more complex approaches in terms of combining cross-sectoral and cross-discipline knowledge, information and data. For example, Steffen et al. (2015) introduce the concept of planetary boundaries to define a safe operating space for humans in the earth system, and thereby using data and models coming from many different domains and background. Such integrated scientific and societal perspectives require the combination of a multitude of data sources and the application of different analytical techniques.

Traditionally, science has operated along disciplinary lines in using and applying its data and analytical tools. Data management and curation was hardly an issue, with data being connected and analysed for separate applications and with researchers working with data files on their own computers and not actively publishing or sharing these. In roughly the period 1985–2005 there was a large focus on developing models for knowledge derivation from available data, see for example a review of farm models in Janssen and Van Ittersum (2007), crop models in Van Ittersum and Donatelli (2003), ecological models in Schmolke et al. (2010), land use models in Verburg et al. (2004). This period was followed in 2000–2012 by a period of building modelling frameworks as a method of combining more comprehensive analysis for decision making (e.g. Argent (2004); Van Ittersum et al. (2008); Van Meijl et al. (2006); Knapen et al. (2013)), combined with many information technology and computational innovations to enable rapid analysis of large amounts of data within a single discipline (e.g. Villa et al. (2009)). As a consequence, at this stage the capabilities within disciplines for data processing and analysis are well developed, just as the high level linkage of models in abstract modelling frameworks, even if the methodological framework underlying such efforts is often lacking (Janssen et al., 2011). Looking at Wang's Levels of Conceptual Interoperability Model (Wang et al., 2009), in environmental modelling and simulation there has been substantial and useful progress at the lower levels of technical and semantic interoperability. To advance, besides further addressing these lower levels, also the still unexplored higher levels of semantic and conceptual interoperability have to be targeted.

Fortunately, in recent years a number of trends have emerged

* Corresponding author.
E-mail addresses: rob.lokers@wur.nl (R. Lokers), rob.knapen@wur.nl (R. Knapen), sander.janssen@wur.nl (S. Janssen), yke.vanranden@wur.nl (Y. van Randen), jacques.jansen@wur.nl (J. Jansen).

http://dx.doi.org/10.1016/j.envsoft.2016.07.017

that could fundamentally change this status over the coming decade. First and for all, the mentioned trend of broadening policy and decision contexts research has challenged the science domain in general towards much more multi-disciplinary and integrative research, while the pace of decision making also puts pressure on the timeliness of research results. Second, the political attention has turned to open data as public good resource, as witnessed by open data initiatives (e.g. Global Open Data for Agriculture and Nutrition, www.godan.info), open data conferences and open data portals (e.g. data.gov, data.gov.co.uk, data.overheid.nl, data.fao.org). This development was preceded by a movement to make scientific publications available as open access, which has led to specialized journals being set up and traditional journals offering the option to publish under open access licences. Third, the amount of data available for science has grown enormously in the past years, driven by technology developments such as open access repositories of remote sensing images, the advance of the mobile phone enabling crowd sourcing and citizen science and digital connectedness through social media and internet of things. Fourth, the computational resources have massively increased over the past decades, according to Moore's Law, with also a better availability and accessibility of storage and computational resources in the cloud such as Platform-as-a-Service (PaaS) and Model-as-a-Service (MaaS) technologies.

These developments of more (open) data and higher connectedness in principle offer opportunities to support larger, faster and more complex data-intensive processing and analysis across disciplines as required for supporting evidence-based decision making towards societal challenges. Against this background, recently Big Data has emerged and to some extent has been hyped as a new trend to provide unlimited capabilities in analysis of data, providing revolutionary new insights (McAfee and Brynjolfsson (2012); Boyd and Crawford (2012); McKinsey Global Institute, McKinsey (2011)). Related to the agro-environmental domain, Vitolo et al. (2015) have investigated web technologies dealing with "Big Environmental Data", while Lokers et al. (2015) explore the use of semantic technologies to improve access to Big Data in agriculture and forestry science. For the purpose of this paper, Big Data is defined as: a term encompassing the use of techniques to capture, process, analyse and visualize potentially large datasets in a reasonable timeframe (as defined by NESSI (2012)), while incorporating both structured and unstructured data and covering several disciplines and domains. This definition primarily focusses on technology and on the technological support of some of the elementary data-intensive tasks in science. Use cases on data management in research (Lokers et al., 2014) show a variety of technological challenges associated for instance with environmental modelling, that range from metadata oriented information retrieval issues to heavily data-oriented problems related to Big Data mining and data integration. These challenges in particular concern the effective discovery of the appropriate data for a specific research task. In data-intensive research areas like agro-environmental modelling we have reached the point where automated procedures for selection, collection and indexing are becoming indispensable to effectively exploit this global network of data.

In this paper we examine and analyse use cases from three European projects as guidance to describe current possibilities and future challenges for deployment of Big Data techniques in the field of agro-environmental research, facilitating decision support at the level of societal challenges. For that purpose, a theoretical framework is proposed that allows positioning of Big Data challenges and techniques in the context of interdisciplinary science and the policy-science interface. This framework is then applied to analyse three scientific cases in the agro-environmental domain and to reflect on the current state of play of the application of Big Data technologies in the domain. Based on the analysis of the cases along the theoretical framework, overall observations are made on technology readiness and suggestions are provided for further developments.

2. Analysis

2.1. Theoretical framework

It is useful to start from a theoretical framework framing the complexity of challenges and demystifying the hype of Big Data. Such a theoretical framework needs to be tailored to the context of the agro-environmental domain. To achieve this, Big Data, its characteristics and ways of processing should be connected to the context of evidence based decision making and to the specifics of data-intensive challenges in the agro-environmental domain.

To frame the way (big) data is used in decision making we introduce a knowledge management model, extending a broadly used and recognized concept which has been elaborated on in numerous publications in different forms and under different names and to which we will refer here as the data–information–knowledge–wisdom or DIKW hierarchy (Rowley, 2007).

The model (see Fig. 1) is used to contextualize data, information, knowledge, and sometimes wisdom, with respect to one another and to identify and describe the processes involved in the transformation of an entity at a lower level in the hierarchy (e.g. data) to an entity at a higher level in the hierarchy (e.g. information). The idea is that decision makers need 'wisdom' for taking evidence based decisions. Such wisdom can be developed by combining available knowledge with less tangible assets like interests, values, preferences, ethics etc. The knowledge base they use is essentially derived from data. Data can in this respect be considered the raw material to produce information through the addition of meaning. Information is again enriched, creating knowledge by using and combining decision and policy contextual applications like for instance integrated models, impact assessments or decision support systems.

Agro-environmental research use cases usually concern dynamic systems with complex interactions between living organisms or perishable products (e.g. plants, animals, humans, agricultural products) and their environment. Describing such systems requires complex and usually detailed information regarding status and behaviour of its entities and their environmental conditions. It can include its actual status, but also historical or predicted future conditions. Because of the spatial dynamics and temporal variability of living systems, data regarding the temporal and spatial behaviour of entities and local conditions are essential. Moreover, understanding these interactions requires the observation, analysis and integration of knowledge of subsystems of very different nature, for example biological, climate, soil and water subsystems. The complexity of describing, analysing and understanding such systems and the magnitude and heterogeneity of the data involved can be easily understood.

The complexity of handling Big Data is highly associated with its typical characteristics, often described as the "3 V's" of Big Data, i.e. *Volume*, *Variety* and *Velocity* (Laney, 2001).

Volume refers to the unprecedented amounts of data becoming available through new technologies supporting massive generation or collection of data and efficient means of storage. Relevant examples for the agro-environmental domain include climate data (especially climate projections) and remote sensing data. Terabyte to Petabyte size volumes are easily reached when attempting to capture - for example - natural variability on detailed spatial and temporal scales.

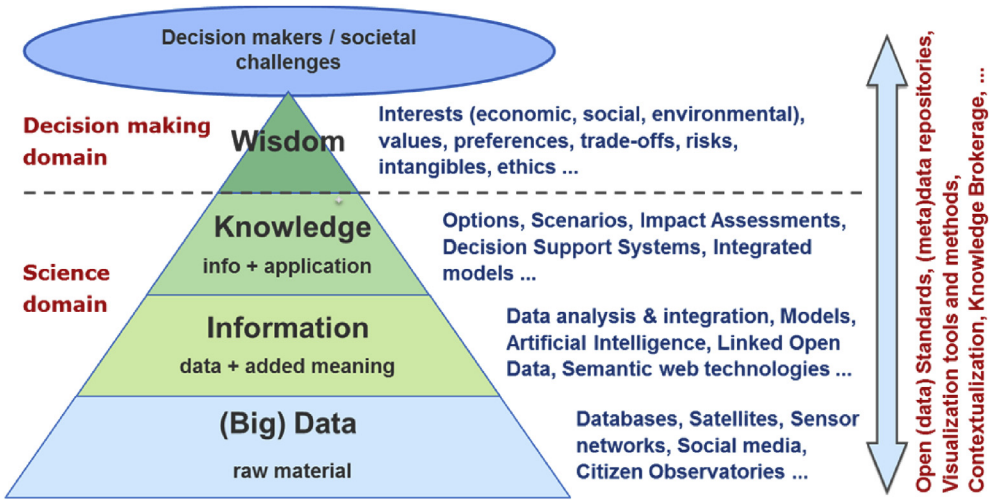

Fig. 1. DIKW hierarchy, from Big Data to decision making for societal challenges.

Velocity refers to the pace at which new data is becoming available, e.g. through real-time data streams, but also refers to the usually high requirements regarding processing time to make the data and its value-add derivatives available for end users. In the agro-environmental domain, real-time data generated by sensor networks or citizen science networks are good examples of such streams, while monitoring and early warning systems commonly require near real-time processing of such data streams in order to provide timely information to decision makers.

Variety concerns the ever increasing heterogeneity of data relevant for decision making. Firstly, this is caused by the continuous evolvement of available streams and formats, e.g. from social media and mobile applications. Moreover, information from an increasing range of disciplines is needed, in particular in the agro-environmental domain. This is due to the many subsystems of very different nature, the tremendous width of current societal challenges to be addressed and the resulting complexity of associated decision contexts. Because individual disciplines tend to have a background of working in silos and using their own tailored data formats and vocabularies, these attempts to integrate data or information from different domains face a multitude of technical and semantic challenges.

In addition to the three V's mentioned, additional characteristics of Big Data have been identified. *Veracity*, often mentioned as being "the fourth V" (http://www.ibmbigdatahub.com/infographic/four-vs-big-data), seems to be the most relevant one when we specifically consider the agro-environmental domain. Veracity, which addresses among others the integrity and accuracy of data and data sources, is highly associated with trust and with having confidence that the quality of data is sufficient to serve as evidence base for critical decision making. Researchers will have to leave the safe environment of familiar data silos in peer networks, while at the same time the growing size and complexity of the data ecosystems grows beyond the capacities of a human being to judge the quality of all associated data sources. Consequently, frameworks and working procedures that ensure integrity of data and its derived products and trustworthy indicators for integrity become indispensable.

Fig. 2 shows how these Big Data characteristics are linked to the DIKW layers when we also consider that in most agro-environmental cases multiple disciplines are involved, with different content regarding data, information and knowledge and different perspectives on policy and decision making.

In the context of Big Data, the DIKW hierarchy also conceptualizes the process of turning the enormous mass of data, which as a raw material has little or no significance to end users, into compact, structured and contextualized, manageable 'chunks' that are applicable in a specific decision making context. End users will implicitly presume that these have been synthesized using the most appropriate sources from the Big Data pool, interpreted and processed according to their decision context, using the most reliable and timely information available. Evidently, such presumptions pose an enormous challenge to the whole community of ICT-experts, data scientists and domain experts that are involved in handling the various steps in this process. The broad scope, both vertically over different ICT, data science and knowledge management expertise areas and horizontally, covering the multi-disciplinary of present-day decision contexts, requires a highly cooperative approach and the establishment of harmonized concerted processes, organized through a combined top-down and bottom-up approach.

To explore the possibilities to meet the challenges described above, in the next section three data-intensive use cases from the agro-environmental domain will be described and analysed with regard to their position in the theoretical framework and the associated Big Data characteristics. Table 1 summarizes the linkage of the cases with the Big Data characteristics and the DIKW model described above.

2.2. Case: semantic driven discovery

2.2.1. Problem statement

This use case addresses the harmonized provision of scattered and heterogeneous data for impact assessment to decision makers and researchers. An impact assessment study typically requires assessing the potential economic, social, and environmental effects of alternative policy options through a number of scientific computer models, which can span various science domains. Each of the models requires sets of trustworthy input data. For example, agricultural impact assessment studies could use scientific models such as: APES - a cropping system model (Donatelli et al., 2010); FSSIM - a bio-economic farm model (Louhichi et al., 2010); CAPRI - an agricultural sector model (Britz et al., 2007); and GTAP - a computable general equilibrium model for global markets (Hertel, 1997). Input data required would include, amongst others, crop parameter data, data on local soil types, historical and simulated

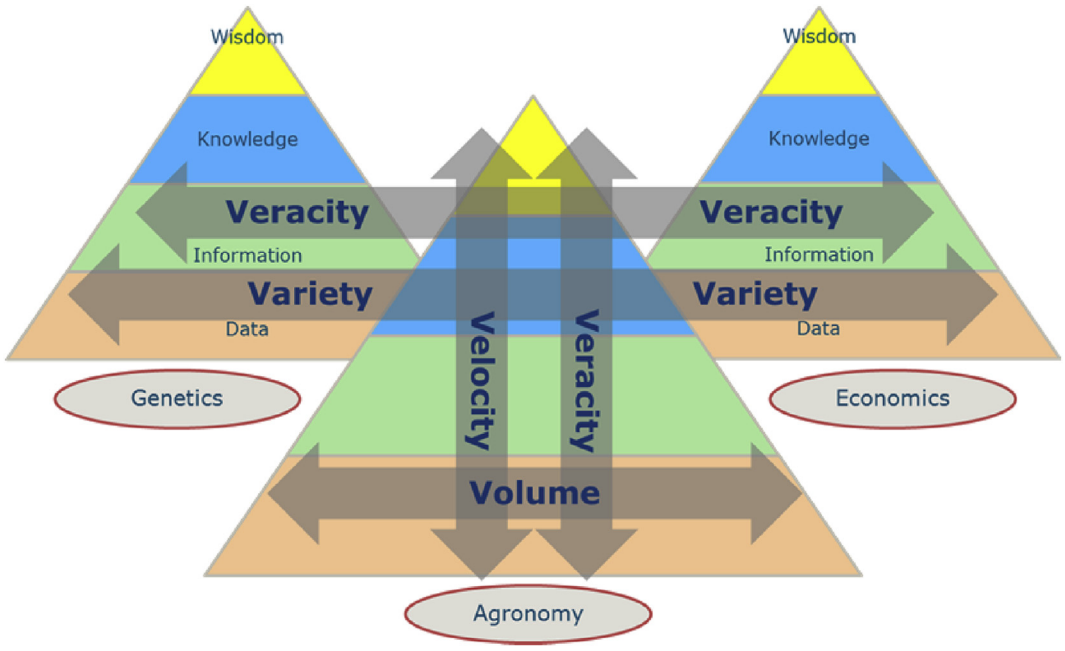

Fig. 2. Multidisciplinary Big Data pool and characteristics.

Table 1
Analysed use cases characteristics.

Use case	Volume	Velocity	Variety	Veracity	D	I	K	W
Semantic driven discovery		X	X	X	X		X	X
Data driven discovery	X		X	X	X	X		
Big Data querying	X	X	X		X	X		

future weather data (on local, regional, and global scale)). Many of such datasets are available, either locally in organization's repositories, or on the Internet as open data. Due to the expanding use of sensors and satellites that can measure e.g. crop, soil and meteorological data at increasingly finer temporal and spatial resolutions, not only the amount of available datasets is growing, but also their sizes. This makes finding the usable pieces of data one of the key challenges of Big Data.

In an approach to address this discoverability challenge, the LIAISE project developed the LIAISE Toolkit (http://www.liaise-kit.eu). LIAISE — Linking Impact Assessment Instruments to Sustainability Expertise — was established in 2009 to improve the application of Impact Assessment (IA) by both the research and the policy making communities. The Toolkit facilitates the categorization and discoverability of metadata for different types of knowledge resources related to IA, for example datasets, scientific models, frameworks, practical examples and domain experts. Submitted information is categorized into topics by key experts before it is published and made accessible through the Toolkit website. Initially this website supported directory based discovery with a faceted search mechanism. Following technological developments near the end of the project it was explored how the search capabilities could be improved by the use of semantic technology. In particular this included investigating whether and how recent Natural Language Processing (NLP) and Machine Learning (ML) techniques could be used to automatically derive required metadata from unstructured text sources and relate it to a defined LIAISE overall ontology. It was assumed that using these techniques could lead to a system which does not only rely on manual provision of metadata by experts, but which can also get its content from automated discovery of relevant metadata, or enriching existing sparse metadata from auxiliary documentation such as reports, published papers, or websites. It would support finding the relevant small pieces of data (at the DWIK Data layer), as well as make the system more “intelligent”, operating at the Knowledge and Wisdom layers, linking available heterogeneous knowledge sources from multiple disciplines to specific contexts of decision and policy making. Furthermore, the case is strongly connected to the variety characteristic, establishing semantic links between the knowledge and decision making layers and by exposing new opportunities for innovative re-using and combining tools in new domains. Through its foreseen approach of automated linkage, it also touches the aspect of improving velocity, while at the same attempting to retain veracity or trust in the generated knowledge.

2.2.2. Methodology and implementation

For its practical implementation, the case aimed at extending the existing LIAISE Toolkit with (i) a way to use the LIAISE ontology for linking existing external datasets to all already available knowledge resources in the Toolkit, and (ii) the use of a similar pathway to relate typical impact assessment questions to relevant knowledge resources in the Toolkit, providing a semantic search mechanism. Fig. 3 illustrates the foreseen steps, including: (1) selecting and processing of auxiliary documentation of simulations models and datasets using NLP techniques, (2) mapping of the data to the defined LIAISE ontology, and (3) storing it. For retrieval through a stand-alone web interface (6) or from the Toolkit website, questions posed in natural language will be processed (4), related to the stored information and used to find (5) matching

search results (i.e. relevant knowledge resources).

As a proof-of-concept the semantic linkage exercise was developed around the datasets provided online by the European Environmental Agency (EEA, http://www.eea.europa.eu/data-and-maps). The metadata of these EEA datasets contains references to a list of topics relevant for EEA resources (see http://www.eea.europa.eu/themes). LIAISE, on the other side, uses a taxonomy of impact areas for tagging included knowledge items.

For reasons of performance and quality an automated procedure periodically retrieves metadata of available EEA datasets from their semantic web SPARQL endpoint. It then attempts to find and add relevant LIAISE taxonomy-based impact area tags to the metadata, thus establishing links that allow the LIAISE web portal to mention the datasets at appropriate places, for example as potentially suitable input to a simulation model. To create the links, the automated procedure needs to perform some kind of semantic comparison based on available metadata and/or data. Several approaches for such a semantic comparison were explored and tested.

The first approach was based on the exploitation of Machine Learning and Natural Language Processing techniques, enabling computers to derive meaning from human or natural language input (Ng and Zelle, 1997). It foresaw the building of a corpus for the automatic determination of relevant terms from the metadata available through LIAISE knowledge resources to subsequently analyse and tag the external metadata with a Machine Learning algorithm. Unfortunately, at the time of development the Toolkit was just started to get filled by the experts and a corpus of adequate size could not yet be constructed within the project time boundaries. Existing resources such as the online, publicly available dictionary WORDNET (Miller, 1995) do not contain the very domain specific knowledge required for e.g. semantic parsing, and a sense-tagged corpus needs to be added to improve automated semantic interpretation. Finding sufficient material for building training data for machine learning was an additional, yet related problem. Consequently, initial ambitions for this case had to be scaled down.

A second, less elaborate approach based on textual matching techniques was subsequently explored. Using the OpenNLP tools (http://opennlp.apache.org), each textual description of a LIAISE taxonomy term for an impact area (e.g. “Environmental Impacts - The environmental consequences of firms and consumers - Sustainable production and consumption”) was syntactically analysed to find all nouns in it, and compared to nouns found in any text field (title, description, topics, etc.) of each resource from the EEA ontology referring to a dataset. The more nouns matched, the more relevant the resource was considered, and the higher it was ranked in the search results. This simple approach proved to be relatively successful due to the fact that both EEA and LIAISE work in the environmental science domain and thus already use a kind of shared vocabulary. Their words and terms in most cases mean the same things. Yet, LIAISE topics and subtopics purposefully have broad and non-restrictive titles so that experts can always find one or more topics their knowledge resources fit in without having to define new topics. While this keeps the taxonomy stable, it makes it harder to use for machine processing. Hence precision and recall of the text matching turned out to be too low to make it an acceptable approach, and the search results contained too much noise over signal to make it acceptable to the users.

Therefore, the final implemented approach was an expert-driven linkage process. This method uses a mapping table in which the expert manually links the LIAISE impact areas (or taxonomy terms) to EEA thematic topics. It does not provide an explicit indication of the quality of the match, which is implicitly associated with the expert and their level of expertise. Because this mapping requires manual input by experts for each data provider to be added to the system and for changes in the taxonomies, it is more time consuming and less dynamic, but it does provide expert based quality of the links, creating trust for the web portal users, thus addressing the veracity aspect.

2.2.3. Results

From the three explored approaches to link (EEA) datasets with the knowledge base available in the LIAISE Toolkit, the semi-manual method where experts manually link Impact Assessment terms to terms related to the datasets was implemented. The two

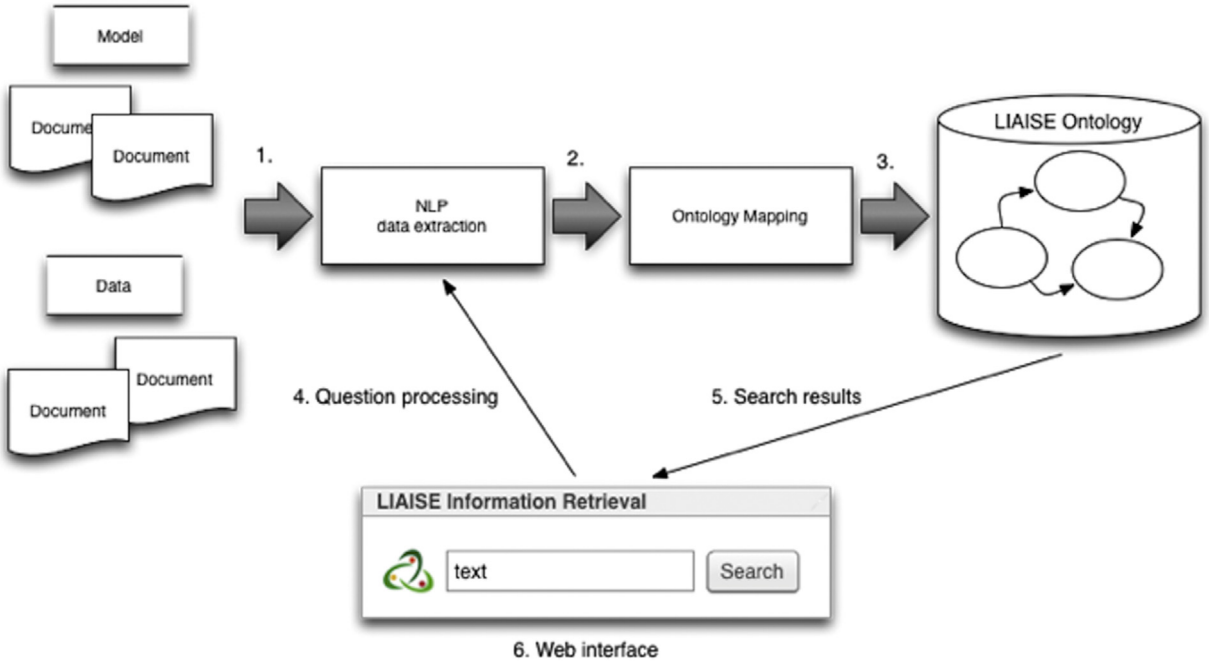

Fig. 3. LIAISE semantics driven information retrieval concept.

automatic processes that were examined proved to be too costly or ineffective in this particular case. For the Machine Learning and NLP-based approach, the main barrier was that building a working corpus for this purpose from available LIAISE resources was not possible, due to the lack of sufficient material available at the time of development, and the unforeseen amount of time that it would take.

Using textual matching, it appeared that in all cases the process resulted in considerable amounts of erroneous matches (low precision and recall values) producing undesired and unusable results. This could be at least partially assigned to the relative simplicity of the non-semantic textual matching techniques used in the process. Moreover, matching also failed because the process operated on extracted high level generic terms (water, air, pollution) instead of more specific compound terms (also known as n-grams) like 'surface water' and 'air pollution'. Retrospectively it can be concluded that trust, or veracity, also plays a relevant role. Even if conditions are met to successfully implement automated procedures for tagging, it will remain hard at the moment for these technologies to gain the level of trust that the scientific user community tends to exhibit if experts perform the job manually.

2.3. Case: data driven discovery

2.3.1. Problem statement

The Trees4Future project (www.trees4future.eu) is an integrative European Research Infrastructure project that aims at integration and further development and improvement of major forest genetics and forestry Research Infrastructures. One of the objectives is to make forestry scientific data discoverable and accessible for a broad audience of modellers and decision makers in and outside the forestry research community. Like in many other scientific domains, forestry researchers traditionally rely on their own peers and scientific networks when collecting the data required for their work. Only recently the forestry research community has started to harmonize and share their data, especially in the area of genetics. However, a lot of relevant data is still stored in silos, sometimes even in local or private repositories. Moreover, datasets often are not documented with appropriate metadata. In many cases, researchers do not see the benefits of documenting data, or data is consciously kept private for example because associated research results are still to be published or because of fear for misuse. In general there is often no incentive, nor sense of urgency to actively share data other than through (trusted) networks and personal contacts. This corresponds to observations in literature, suggesting that apart from the technology challenges, many disciplines also still lack the institutional and cultural frameworks required for efficient data sharing, together leading to a "scandalous shortfall" in the sharing of data by researchers ("Data's Shameful Neglect [Editorial]" (2009)). Thus, valuable research data is hard to find without knowing the right people, and only partially available for the whole community of interest. Consequently it still remains hard to acquire the specific targeted data for interdisciplinary work. This lack of discoverability is an even more pressing issue for "newcomers", for scientists from associated domains that require forestry data for their work or for decision makers looking for evidence based information.

One of the research communities in the Trees4Future project are forestry modellers. Their work on present-day societal challenges (e.g. related to bio economy, climate change) requires interdisciplinary approaches, like integrated modelling. As an example, assessing climate change impacts and exploring climate adaptation strategies requires coupling of models that describe various subdomains and cover different spatial and temporal resolutions. In Trees4Future, such integrated assessments required the linkage of the ForGEM model (Kramer et al., 2013), the EFISCEN model (Nabuurs et al., 2000) and the Tosia model (Lindner et al., 2010). While the ForGEM model assesses genetic adaptive responses on the individual tree and population level, the EFISCEN model projects forest resource development on a regional and European scale, and the Tosia model analyses environmental, economic, and social impacts of changes in forestry-wood production chains. Heterogeneous data, varying from detailed genetic data, phenotypic traits and high resolution climate and soil data, to statistical data on species distribution, forest management practices and market information are required to address such integrated modelling exercises. Given the current disconnectedness and lack of context, it is quite complex and time consuming to discover and get access to these data. The Trees4Future project aims at improving this situation by developing technical solutions to facilitate the documentation, publication and discoverability of forestry data by setting up a forestry data infrastructure. Moreover, through this infrastructure it aims at demonstrating benefits and fostering broader uptake of data sharing and documentation practices.

From the perspective of the theoretical framework, this case is strongly connected to veracity, addressing issues of trust and quality. This obviously works in two directions. On the data owner side, there needs to be trust that data is sufficiently documented and will not be misinterpreted or misused. Data consumers, on the other side, should have trust in the reliability and correctness and completeness of associated metadata. The case also addresses variety, through the requirement to provide integrated access to sources coming from a range of relevant subdomains and the related need to provide semantic linkages over the associated (meta)data. This case mainly concerns the lower levels of the DIKW hierarchy (the data and information level) and the need to make the available data and information part of the multidisciplinary Big Data pool, adding the required context on the dataset level to make data potentially usable for the broader community.

2.3.2. Methodology and implementation

To improve access to data required for and generated by forestry research, a data search and discovery service on top of a federated metadata repository was developed. Main objectives were firstly that the system had to be able to provide both already documented and accessible datasets and up till now inaccessible datasets, thus connecting not only organisations that have already organized and standardized their processes but also the smaller organisations and individuals that are not equipped with the required infrastructure. Secondly, end users were to be provided with a facility to easily search and discover available forestry data, once datasets have been documented and published through the developed mechanisms. This search function was considered to be the necessary "proof of the pudding", required to convince end users to use the system, but also to convince data owners of the benefits of documenting and publishing their data.

To achieve improved discoverability, the following components were developed to support the data publication process depicted in Fig. 4:

- A concise metadata schema, based on the widely supported and extensible Dublin Core standard and extended with additional metadata elements to support forestry specific metadata;
- A public metadata registry, composed of an online metadata editor and an underlying repository, which publishes its metadata records through the OAI-PMH protocol, providing a standardized and harvestable metadata endpoint;
- A forestry ontology that allows the conceptualization of datasets and its interlinkage with commonly used external ontologies (e.g. AGROVOC (FAO, 2016), a genetic traits ontology);

- A metadata harvesting, triplification and annotation mechanism that supports harvesting metadata following the developed forestry metadata schema as well as standardized metadata schema's like INSPIRE and ISO; decomposes the metadata into ontology concepts using among others natural language processing (NLP) techniques and stores these in an RDF (Resource Description Framework) database; and links the derived dataset concepts to the concepts of the available external ontologies;
- A semantic search mechanism and search interface, allowing users to transparently search the registered datasets, using the power of semantics in the underlying RDF store.

2.3.3. Results

The developed infrastructure clearly has increased the discoverability of forestry research data and improved its availability for a broader audience. It covers the variety of data required for integrated forestry modelling cases, like the described climate adaptation use case and others. This is, first of all, because it provides federated access to the currently scattered and sometimes inaccessible wealth of forestry research data. The developed data infrastructure already publishes metadata of more than 300 datasets from major European data repositories, and offers the option for small organisations and individuals to publish their (meta)data through a managed access point. Moreover, it offers opportunities to publish reference datasets for integrated modelling, providing less experienced modellers with entry points to build their experiments.

The developed infrastructure has been tested with a set of queries to evaluate its added value through the use of semantics. Typical examples include the linkage of synonyms (e.g. rainfall results in datasets tagged with the concept rain) and broader and narrower terms (e.g. precipitation results in datasets tagged with rain, snow, hail). These tests, and the first impressions of its use in practice, show that even with relatively simple knowledge technology additions, well documented data can be made accessible in a better way, making it easier to discover data through better structuring, indexing and search capabilities. The addition of semantic capabilities and its ability to directly search topics, concepts and associations linked to a vast number of sources to a metadata repository increases the discoverability of datasets, because it reveals otherwise unknown linkages between the common vocabularies of different users and the actual metadata concepts. This is accomplished by (1) returning results that are semantically related to the provided search terms and (2) revealing related terms to the user when composing their search conditions (e.g. by a semantically driven autocomplete function). An observed additional benefit is that providing the scientific community with improved discovery mechanisms increases awareness that data documentation is important, and contributes to insights in how data could be best documented in order to provide added value to end users.

On the other hand, we also conclude that in forestry and related domains the currently available metadata is scarce and often of low quality, which complicates the linkage of metadata concepts with (external) ontology concepts. A second observation is that currently available metadata standards provide insufficient possibilities to (automatically) select datasets that fit researchers needs, e.g. in technical domains like modelling and simulation. In general metadata schemas lack the structure and depth required to structurally capture the complexity of scientific datasets. Commonly used and essential fields, like for example lineage, do not provide the structure required to address the complex production processes of data. Moreover, the lack of depth prevents that the structure and contents of the data itself (for example its attributes and datatypes) can be addressed in a structured manner. In the use case, this issue was tackled by combining and linking isolated fragments of a broad coverage vocabulary (AGROVOC) with specific and detailed subdomain specific semantics. Obviously, this is a very customized and elaborate approach and not a viable generic solution.

2.4. Case: Big Data querying

2.4.1. Problem statement

Research in the agro-environmental domain has to deal with large and very diverse datasets, both in content, structure, and storage format. Because of the current move towards open access and open data, an increasing amount of data is brought out of their information silos and made accessible as part of what is called the

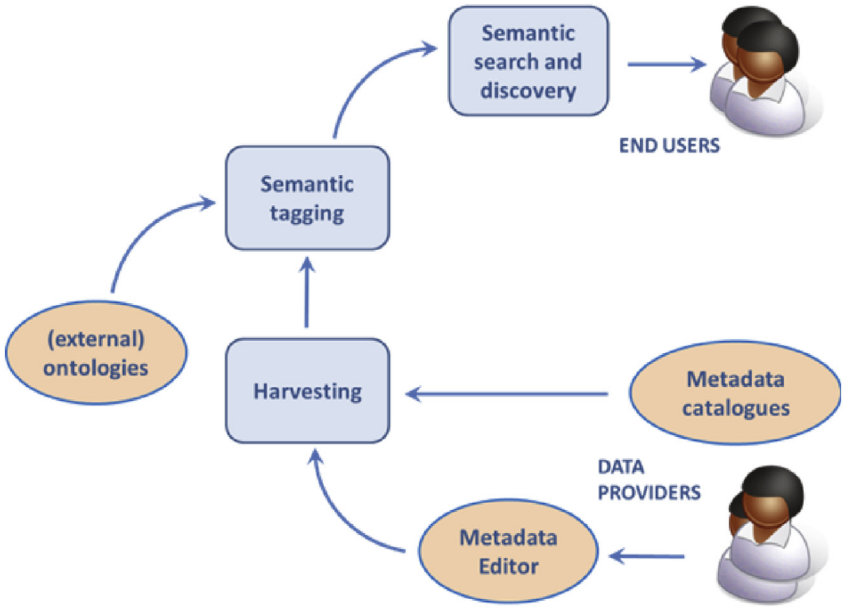

Fig. 4. Publication process workflow developed in Trees4Future.

Linked (Open) Data (LOD) cloud, resulting in an extensive network of distributed heterogeneous data sources. Unfortunately, access to this network to date is neither easy nor transparent, and current centrally-managed or even distributed data repositories are not able to meet the data science challenges ahead, starting with adequate Big Data querying facilities. The EU FP7 research project SemaGrow examined solutions to provide more effective and transparent ways to access distributed data. It aimed at developing algorithms and infrastructure for the efficient querying of large-scale federations of independently-managed data sources, i.e., the nodes of the Linked Data cloud. To address the differences in storage formats, it builds upon the already established and frequently used principles behind the Semantic Web, namely RDF and the SPARQL query language. These standards enable the sharing and reusing of data across applications and scientific community boundaries, and allow the interconnecting of data in the LOD cloud.

SemaGrow specifically focussed on the agriculture domain and its use cases through a series of data pilots, exploring the specific data challenges of this domain. These challenges typically include discovery, merging and integration of large and very diverse spatio-temporal datasets. One of the use cases explored in SemaGrow is regional agro-climatic modelling in the frame of climate adaptation. Climate parameters required for regional modelling are usually stored in large multidimensional files, often with global or transregional spatial coverage and long term temporal coverage. Modellers tend to duplicate large amounts of data for their modelling experiments, which are then locally processed to the required extent and scales. Besides the general issue of resource efficiency, such ways of working can pose significant barriers, specifically in regions where networking, storage and computing resources are limited. SemaGrow has examined ways to allow the thematic, spatial and temporal querying and merging of large distributed datasets, returning relatively light and integrated datasets. As an example, this would allow an agricultural modeller in Ghana with limited networking and storage resources to acquire a merged subset of temperature, precipitation and soil parameters for a specific region in the country.

With regards to the theoretical framework, the case primarily focussed on the volume and variety characteristics of Big Data, exploring ways to allow scientists to efficiently access large, distributed data sources in a federated manner and to download and merge manageable subsets of different nature. Although velocity was not the primary focus, it is relevant to note that the case elaborates on automating data integration problems that generally are very labour and time intensive through the involvement of experts of different disciplines. While developed solutions might technically be considered as non-performant, they could still result in dramatic improvement of efficiency in the face of timely provision of information required for decision making. It concerns mainly the data and information layer of the DIKW hierarchy, attempting to efficiently bridge the gap between these levels by automatically processing and harmonising sources from the Big Data pool to a level that offers better opportunities for connecting data with the tools (e.g. models, data analysis) operated on the information level. Consequently, it not only potentially reduces the efforts and resources required to produce information from raw data, but also touches some of the integration challenges associated with interdisciplinary science.

2.4.2. Methodology and implementation

To be able to demonstrate SemaGrow Big Data querying capabilities in the frame of real-world applications, and to be able to compare its characteristics to a reference situation, as one of the pilots the Trees4Future Clearinghouse system described in the previous case was adapted to work using SemaGrow technologies. For that purpose, the Trees4Future back-end was replaced with the infrastructure developed by SemaGrow, the so-called SemaGrow Stack, and a set of distributed RDF databases containing triplified data and metadata. As a result, the demonstrator application also extends the reference application by not only offering the option to perform semantic queries on metadata but also on the underlying data.

The SemaGrow Stack (https://github.com/semagrow/semagrow) is a 'federated SPARQL query processor' that can efficiently query a set of distributed heterogeneous data nodes. It includes a query planner that uses metadata about the nodes of the federation to optimize the query execution. This metadata follows the Sevod vocabulary (http://www.w3.org/2015/03/sevod), also developed in the project, that extends the VoID vocabulary with statistical information akin to database histograms. The Stack uses the reactive software paradigm to properly handle unresponsive or slow data nodes in the federation. As such, the SemaGrow Stack provides a unifying endpoint that allows transparent querying of the underlying triple stores without having to know their (possibly) heterogeneous schemas.

Triple stores were set up, holding triplified agro-environmental data, selected from the ISI-MIP and AgMIP data harmonisation initiatives. ISI-MIP, The Inter-Sectorial Impact Model Intercomparison Project, is a community-driven modelling effort bringing together impact models across sectors and scales to create consistent and comprehensive projections of the impacts of different levels of global warming. Input and output data from ISI-MIP is made available as NetCDF files using the Climate and Forecast conventions for its metadata. The Agricultural Model Intercomparison and Improvement Project, AgMIP, is a major international effort, linking the climate, crop, and economic modelling communities with cutting-edge information technology, to produce improved crop and economic models and the next generation of climate impact projections for the agricultural sector. AgMIP provides data in JSON format using the ICASA Variable List for its metadata. Both data collections are harmonized, but are quite different in nature, e.g. global gridded time-series data of simulation model projections, versus single point location based time-series of field management and weather station observed data. These features make them well suited to evaluate how the SemaGrow Stack handles the heterogeneity related aspects. Data from these sources has been 'triplified' into triple stores, so they can be queried using SPARQL. Besides, the different vocabularies used (CF Conventions for ISI-MIP data and ICASA for AgMIP data) have been aligned through the use of the AGROVOC thesaurus. The amount of datasets that have been triplified for the demonstrator is limited. It concerns around 10 global coverage, long-term ISI-MIP datasets and a few dozen of AgMIP datasets. However, especially due to the volume of the ISI-MIP datasets, the total size was at the Tera triples level, allowing to also explore the volume related aspects and the associated behaviour of the SemaGrow infrastructure.

Lastly a spatio-temporal triple store (Strabon, http://strabon.di.uoa.gr) has been added to the federated nodes so that spatial queries, e.g. point-in-polygon, can be resolved. To connect the web application front-end of the demonstrator with the SemaGrow Stack instance, a small additional layer of middleware software was needed. It translates URL requests including parameter values into the proper SPARQL queries for the Stack, and vice versa pre-processes the raw query results into a response the demonstrator can handle. Furthermore, it is able to create valid NetCDF files from the RDF Data Cube format used internally, to better serve end-users needs.

2.4.3. Results

So far, the described demonstrator application has been tested

by a limited group of end-users. The demonstrator gets positive remarks for the functionality it offers, but people expect better performance both for metadata searches (less than 10 s expected, versus 5–30 s measured) and for data downloads (less than 30 min expected, versus several minutes to several days measured, depending on the size of the selected data), as well as access to much more data. Both can possibly be met by massive upscaling of the infrastructure. Notably, in the performed expert enquiries several experts have explicitly mentioned that, even with the measured response times, the demonstrated querying and data fusion facilities can be quite useful. It should be realized that, for example in the formerly mentioned use case of agricultural modelling in Ghana, composing a dataset for modelling requires different processing steps. It usually requires consultation of, and cooperation with local and remote specialists and consequently aggregated time investments and resulting lead times can be high. Thus, automated data fusion queries, even when taking hours or days, could make the research process in such cases more efficient.

The SemaGrow project has also shown how time-consuming it remains to process data so that it is properly annotated with metadata, triplified, and aligned to make it part of the LOD. While tools for ontology matching and alignment were available through the project, these could not be used because reference vocabularies did not comply with the supported standards. Moreover, it appeared that a commonly used vocabulary like AGROVOC is not well suited to effectively annotate datasets on the level of detail required for the research problems examined. AGROVOC provides relatively rough concepts for specific variables and provides no specific unit taxonomy. However, fitting selections for specific modelling experiments would require more specific specifications to describe, for example, the parameter “mean daily temperature 2 m above ground level” as well as its specific unit of measurement. Besides, in contrast to for example bibliographic data or text documents, the multi-dimensional data used in agro-environmental science still challenges state-of-the-art triple stores and current semantic web technology. Issues like different spatial projections, spatial and temporal scales, unit conversions, handling streaming data or simple data manipulations, could not be considered within the scope of the project, but were on the list of evaluation comments by the end-users. Consequently, providing transparent and unified access to these datasets is not yet trivial.

3. Conclusions and recommendations

Three use cases are described that have addressed different issues related to Big Data usage and technologies in the agro-environmental domain. These have also been put into the perspective of a theoretical framework to structure their complexity. In the analysed cases, a variety of issues were encountered spanning the whole range of Big Data characteristics (the 4 V's) and the layers of the DIKW hierarchy. Cases generally focused on discovering and combining heterogeneous datasets for modelling and decision making in interdisciplinary domains. While it is obvious that challenges regarding the volume and velocity aspects exist, and there are not yet clear solutions in all cases, the contours of future technical solutions are already visible, combining cloud based storage and computing with improved and better integrated infrastructural components. Research initiatives explore and develop innovative infrastructures and several commercial services are offered. More important for the agro-environmental domain is that steps are being taken to improve the handling of Big Data, including the aspect of dealing with the spatio-temporal data that is very common to the domain. Specific processing requirements of this type of data include spatial and temporal up- or downscaling and handling a large variety of spatial reference systems. Such processing could be more effectively handled by an additional software layer, e.g. through a data centric design (http://research.ibm.com/articles/datacentricdesign/), where much of the processing is moved to the places where the data is stored.

More persistent barriers in agro-environmental science, and probably also in other areas that require highly interdisciplinary knowledge for decision making, lie in handling the variety and veracity aspects. Not surprisingly, these aspects are also crucial to link the different levels of the DIKW hierarchy, both vertically, allowing to work up raw data to knowledge fit for decision making, and horizontally, to meaningfully connect content from different disciplines that are currently often disconnected. In order to be able to meaningfully link heterogeneous sources coming from different disciplines and being generated for different purposes, improved semantic interoperability is needed. Possibly it also is needed to strive towards a higher level of the Conceptual Interoperability Model introduced by Wang et al. (2009) (Wang et al., 2009). Based on the work done on the three presented cases, two approaches can be recognized, one top-down driven and the other one bottom-up.

The top-down approach would include defining and agreeing upon a top-level ontology for the agro-environmental domain, and all subdomains relating their specific ontologies to this top-level ontology by harmonizing or alignment efforts. Many vocabularies and ontologies exist in the agro-environmental domain, developed with different purposes and covering different subdomains. They vary from broad coverage and relatively global (e.g. AGROVOC) to very specific coverage and detailed. The scope of agro-environmental scientific challenges usually requires dealing with different vocabularies that are typically not interoperable and sometimes even conflicting, which in practice makes it very hard to align semantics in such a way that the result remains meaningful and is fit for a specific purpose. Therefore, besides the elements of coverage and granularity, serious barriers are the fact that different standards are used, and that alignment is very labour intensive and requires interdisciplinary expertise. The analysed cases particularly show that standardized and broadly accepted semantics to describe datasets on the level of its attributes are generally lacking, and that available ontologies and vocabularies cannot easily be applied. Solutions like combining (fragments of) different semantic sources or manual linkage of vocabularies for very specific purposes can work for the specific case, but are obviously not sustainable. Yet, these, often small, semantic differences between simultaneously existing ontologies competing for adherents may simply continue to exist as part of our academic and political freedoms. Still it would be worthwhile to at least work towards common, linked ontologies instead of ending up with, exaggerating, one ontology per disciplinary data silo.

A more bottom-up oriented approach revolves around the use of semantic interpretation technologies such as Natural Language Processing and Machine Learning algorithms. With more data, including e.g. text documents and web pages, and metadata becoming available, it will become impossible for humans to properly relate the data to ontologies, next to discussions about which ontologies to use. It certainly seems more practical when computers can tag data on an ad-hoc, case-by-case basis, based on some level of understanding of the meaning of the data. The WORDNET dictionary already provides a good starting point, but needs to be extended with domain specific knowledge, like discussed in the use case on semantic driven discovery. Building up such a corpus covering the agro-environmental domain is a time consuming activity, but it would be highly reusable in many future applications. As of now, use of semantic technologies is not very well developed in the agro-environmental domain. Consequently, consistency of produced results is still varying, making its

introduction and acceptance ('veracity') a challenge.

The sketched approaches are of course not mutually exclusive, and might meet somewhere in the middle. A relevant initiative that demonstrates a possible way forward is CYNERGI http://earthcube.org/group/cinergi). This initiative tries to combine bottom-up (e.g. enhancers using semantic techniques to improve metadata) and top-down (e.g. using a generic metadata schema) aspects to harmonize access to interdisciplinary datasets. More importantly, CINERGY recognizes the shortcomings of available metadata, semantics and available technologies, like machine learning and NLP. It explicitly includes human engagement as an indispensable factor in the process of making data fit for interdisciplinary science. Direct involvement of scientists to select relevant data sources, metadata elements and to validate generated metadata and query results is regarded a crucial element to serve cross-domain, fit-for-use data to scientists. This corresponds with the experiences and outcomes of the analysed cases.

Looking from the perspective of the analysed cases at the lower levels of the DIKW hierarchy, the provision of sufficient, high quality metadata hinders the smooth access to and linkage of scientific data sources. To be able to work with the available metadata and to deal with its shortcomings, all observed cases were somehow confronted with the need to develop customized solutions. Applied solutions vary from implementing manual ontology alignment as an alternative for metadata driven automatic annotation to the improvement of awareness and provision of metadata creation and editing facilities. In general, despite the availability of standardized metadata schemas, well documented scientific data is still scarce. This also appears to be a cultural issue, where scientific practice is often still based on working in silos and interchanging among trusted peers, and data management policy is not well developed. There is a clear link to the veracity aspect of Big Data here. Data users need to trust the provided data, which is most clearly expressed by the quality of its metadata. On the other hand, data providers require trust that their data is used in a proper way, which again can be promoted by adequately documented datasets.

Promising and viable approaches for ICT driven mechanisms to improve interdisciplinary data-intensive research using technologies related to the Big Data domain have been identified, examined and implemented in the analysed use cases. Although in most cases implementation was successful, we can also conclude that effectiveness is limited, due to the current state of data management and semantic coverage in the agro-environmental domain. Based on the analysed cases and the above stated conclusions, we recommend that Big Data research, and especially the efforts to be delivered in this area from the agro-environmental domain (in contrast to the more technically oriented ICT research), focusses on variety and veracity challenges. This focus should lead to the improvement of conditions and development and application of methodologies and techniques required to efficiently provide access to and semantically interlink sources from different disciplines. Obviously, this does not only require technological advances, but also a disruptive change of culture and behaviour. To this end, the development and promotion of working demonstration cases in research environments has proven to be a valuable instrument to create awareness and catalyse such change.

Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007–2013) under grant agreements No. 243826. (LIAISE), 318497 (SemaGrow) and 284181 (Trees4Future) and the Ministry of Economic Affairs of the Netherlands.

References

Argent, R.M., 2004. An overview of model integration for environmental applications–components, frameworks and semantics. Environ. Model. Softw. 19 (3), 219–234.

Boyd, D., Crawford, K., 2012. Critical questions for big data. Inf. Commun. Soc. 15 (5), 662–679.

Britz, W., Heckelei, T., Kempen, M., 2007. Description of the CAPRI Modelling System. Final Report of the CAPRI-dynaspat Project. Institute for Food and Resource Economics, University of Bonn, Bonn, Germany.

Data's shameful neglect, 2009. Nature 461 (7261), 145–145.

Donatelli, M., Russell, G., Rizzoli, A.E., Acutis, M., Adam, M., Athanasiadis, I.N., Balderacchi, M., Bechini, L., Belhouchette, H., Bellocchi, G., Bergez, J.-E., Botta, M., Braudeau, E., Bregaglio, S., Carlini, L., Casellas, E., Celette, F., Ceotto, E., Charron-Moirez, M.H., Confalonieri, R., Corbeels, M., Criscuolo, L., Cruz, P., di Guardo, A., Ditto, D., Dupraz, C., Duru, M., Fiorani, D., Gentile, A., Ewert, F., Gary, C., Habyarimana, E., Jouany, C., Kansou, K., Knapen, R., Filippi, G.L., Leffelaar, P.A., Manici, L., Martin, G., Martin, P., Meuter, E., Mugueta, N., Mulia, R., van Noordwijk, M., Oomen, R., Rosenmund, A., Rossi, V., Salinari, F., Serrano, A., Sorce, A., Vincent, G., Theau, J.-P., Thérond, O., Trevisan, M., Trevisiol, P., van Evert, F.K., Wallach, D., Wery, J., Zerourou, A., 2010. A component-based framework for simulating agricultural production and externalities. In: Brouwer, M.F., Ittersum, K.M. (Eds.), Environmental and Agricultural Modelling: Integrated Approaches for Policy Impact Assessment. Springer Netherlands, Dordrecht, pp. 63–108.

FAO, 2016. AGROVOC Multilingual agricultural thesaurus. http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus.

Hertel, T.W. (Ed.), 1997. Global Trade Analysis: Modeling and Applications. Cambridge University Press, Massachutes, USA.

Janssen, S., Van Ittersum, M.K., 2007. Assessing farm innovations and responses to policies: a review of bio-economic farm models. Agric. Syst. 94 (3), 622–636.

Janssen, S., Athanasiadis, I.N., Bezlepkina, I., Knapen, R., Li, H., Domínguez, I.P., Rizzoli, A.E., van Ittersum, M.K., 2011. Linking models for assessing agricultural land use change. Comput. Electron. Agric. 76 (2), 148–160.

Knapen, R., Janssen, S., Roosenschoon, O., Verweij, P., de Winter, W., Uiterwijk, M., Wien, J.-E., 2013. Evaluating OpenMI as a model integration platform across disciplines. Environ. Model. Softw. 39 (0), 274–282.

Kramer, K., Hengeveld, G.M., Schelhaas, M.J., Werf, D.C.v.d., Winter, W.d, 2013. Genetic Adaptive Response: Missing Issue in Climate Change Assessment Studies.

Laney, D., 2001. 3D data management: controlling data volume, velocity and variety. In: Gartner (Ed.), Gartner.

Lindner, M., Suominen, T., Palosuo, T., Garcia-Gonzalo, J., Verweij, P., Zudin, S., Päivinen, R., 2010. ToSIA—A tool for sustainability impact assessment of forest-wood-chains. Ecol. Model. 221 (18), 2197–2205.

Lokers, R., Konstantopoulos, S., Stellato, A., Knapen, R., Janssen, S., 2014. Designing innovative linked open data and semantic technologies for agro-environmental modelling. In: International Environmental Modelling and Software Society (IEMSs) 7th Intl. Congress on Env. Modelling and Software, San Diego, CA, USA.

Lokers, R., van Randen, Y., Knapen, R., Gaubitzer, S., Zudin, S., Janssen, S., 2015. Improving Access to Big Data in Agriculture and Forestry Using Semantic Technologies. Metadata and Semantics Research. Springer International Publishing, pp. 369–380.

Louhichi, K., Kanellopoulos, A., Janssen, S., Flichman, G., Blanco, M., Hengsdijk, H., Heckelei, T., Berentsen, P., Lansink, A.O., Ittersum, M.V., 2010. FSSIM, a bio-economic farm model for simulating the response of EU farming systems to agricultural and environmental policies. Agric. Syst. 103 (8), 585–597.

McAfee, A., Brynjolfsson, E., 2012. Big data: the management revolution. Harv. Bus. Rev. 61–68.

McKinsey, 2011. Big Data: The next frontier for innovation, competition and productivity. McKinsey Global Insititute.

Nabuurs, G.-J., Schelhaas, M.-J., Pussinen, A., 2000. Validation of the european forest in-formation scenario model (EFISCEN) and a projection of finnish forests. Silva Fenn. 34 (2), 167À179.

NESSI, 2012. Big Data: a New World of Opportunities. NESSI White Paper: NESSI.

Ng, H.T., Zelle, J., 1997. Corpus-based approaches to semantic interpretation in natural language processing. AI Mag. 18 (4), 45–64.

Rowley, J., 2007. The wisdom hierarchy: representations of the DIKW hierarchy. J. Inf. Sci. 33 (2), 163–180.

Schmolke, A., Thorbek, P., DeAngelis, D.L., Grimm, V., 2010. Ecological models supporting environmental decision making: a strategy for the future. Trends Ecol. Evol. 25 (8), 479–486.

Steffen, W., Richardson, K., Rockström, J., Cornell, S.E., Fetzer, I., Bennett, E.M., Biggs, R., Carpenter, S.R., de Vries, W., de Wit, C.A., Folke, C., Gerten, D., Heinke, J., Mace, G.M., Persson, L.M., Ramanathan, V., Reyers, B., Sörlin, S., 2015. Planetary boundaries: guiding human development on a changing planet. Science 347 (6223).

Van Ittersum, M.K., Donatelli, M., 2003. Special issue of the european journal of agronomy: modelling cropping systems. Eur. J. Agron. 18 (3–4), 187–394.

Van Ittersum, M.K., Ewert, F., Heckelei, T., Wery, J., Alkan Olsson, J., Andersen, E., Bezlepkina, I., Brouwer, F., Donatelli, M., Flichman, G., Olsson, L., Rizzoli, A., van der Wal, T., Wien, J.-E., Wolf, J., 2008. Integrated assessment of agricultural systems- a component based framework for the European Union (SEAMLESS). Agric. Syst. 96, 150–165.

Van Meijl, H., van Rheenen, T., Tabeau, A., Eickhout, B., 2006. The impact of different

policy environments on agricultural land use in Europe. Agriculture. Ecosyst. Environ. 114 (1), 21–38.
Verburg, P.H., Schot, P.P., Dijst, M.J., Veldkamp, A., 2004. Land use change modelling: current practice and research priorities. GeoJournal 61 (4), 309–324.
Villa, F., Athanasiadis, I.N., Rizzoli, A.E., 2009. Modelling with knowledge: a review of emerging semantic approaches to environmental modelling. Environ. Model. Softw. 24 (5), 577–587.
Vitolo, C., Elkhatib, Y., Reusser, D., Macleod, C.J.A., Buytaert, W., 2015. Web technologies for environmental big data. Environ. Model. Softw. 63, 185–198.
Wang, W., Tolk, A., Wang, W., 2009. The levels of conceptual interoperability model: applying systems engineering principles to M&S. In: Proceedings of the 2009 Spring Simulation Multiconference. Society for Computer Simulation International: San Diego, California.