International Environmental Modelling and Software Society (iEMSs)
2012 International Congress on Environmental Modelling and Software
Managing Resources of a Limited Planet, Sixth Biennial Meeting, Leipzig, Germany
R. Seppelt, A.A. Voinov, S. Lange, D. Bankamp (Eds.)
http://www.iemss.org/society/index.php/iemss-2012-proceedings

A generic data schema for crop experiment data in food security research

Sander Janssen[a], Daniel van Kraalingen[a], Hendrik Boogaard[a], Allard de Wit[a], Jappe Franke[a], Cheryl Porter[b], Ioannis N. Athanasiadis[c]
[a] *Alterra, Wageningen UR sander.janssen/ Daniel.vankraalingen/ Hendrik.boogaard/ allard.dewit/jappe.franke@wur.nl*
[b] *University of Florida, cporter@ufl.edu*
[c] *Democritus University of Xanthi, ioannis@athanasiadis.info*

Abstract: In agricultural research targeted at food security, crop experiments in fields are a crucial source of information for statistical or model based analyses or purely a system description. In these crop experiments or field trials, crop responses are investigated to a change a management or in different climatic or soil conditions, and thus provide an understanding of production potential in different circumstances. Though crucial, these crop experiments are currently poorly available to the crop research community, which proves an obstacle to developments in the domain. The aim of this paper is to propose a generic data schema, Spatial Temporal Attribute Catalogue, that can be used to store data on agricultural systems compiled with many different purposes and scopes. The generic data schema covers aspects of soil, climate, location, crop management and crop variety characteristics. The data schema is developed in a context of different ongoing and past efforts in structuring this crop experiment data, e.g. the AgMIP crop experiment database, the Global Yield Gap Atlas, and the MOCASSIN project on winterkill. Future developments on the data schema include assessing the possibilities to broaden it to different domains (i.e. socio-economic, ecology, and animal sciences) and the use of semantic technologies for storage and availability.

***Keywords**:* data schema, agricultural systems, generic, model-based assessments

1 INTRODUCTION

The current agricultural scientific community (agronomy, livestock research, agro-economics) is fragmented in its data management, with each project and institute generally realizing its own solution, and a lack of commonly available reference data on (benchmarking of) agricultural production (White et al., 2008; Evert et al, 1999; Janssen, et al. 2009). A clear necessity was signalled for coordinated improvement in data management in several international scientific communities, institutes and development agencies, e.g. in Agricultural Model Improvement Project (www.agmip.org; positioning note), in a database meeting in Dubai for several Bill and Melinda Gates Foundation funded projects and in the IFPRI-lead Geoshare project initiative (www.geoshareproject.org).
In the past, this problem of fragmentation and lack of data availability was already tackled in various initiatives, unfortunately not with conclusive results. For example, as part of ICASA (International Consortium for Agricultural Systems Applications, www.icasa.net), an effort was made to develop standards for documentation and storage of agricultural experiments (Hunt, et al., 2006), and developing exchange mechanisms through a joint portal of such data for scientists. Evert, et al. (1999a, 1999b) proposed a shared schematization of input and output data to cropping systems analysis, which was not taken up afterwards. At the same time, contrarily,

in soil science significant progress was made with storing and delivering soil-related data, at different spatial levels (nationally, Netherlands: www.bodemdata.nl, European: http://eusoils.jrc.ec.europa.eu/data.html, globally: www.isric.org/data/wosis).

Lately, there is a renewed interest on improving the availability and overcoming the fragmentation in agricultural research data. Two important examples are led by international institutes. First, agtrials.org is an initiative of CGIAR institutes through Climate Change, Agriculture and Food Security program to make agricultural trial data available at CGIAR institutes publicly available to researchers around the globe. At the moment, it focuses on making the description of the trial (i.e. the meta-data) available, while not the data itself. Second, FAO has the Agricultural Information Management Standards (aims.fao.org), with the agrovoc ontologies at its core, which concentrate on providing joint meta-data and shared conceptualizations for agricultural research, while not considering the data itself. In a research based effort, agricultural systems data for Europe has been consistently integrated in a joint database through typologies, spatial framework, and a joint conceptualization in an ontology (Janssen, et al., 2009).

Many of the more recent efforts are somehow influenced by a trend towards open linked data in the software engineering and library science domains, which is a trend towards open access and relating or linking different data sources in the public domain, or collected with public resources (Berners-Lee, 2009; Bizer et al. 2012). Agricultural research data at international institutes was available to some extent, but most institutes reviewed had a limited available and documentation of their data (Besemer, et al. 2011).

With the developments towards open linked data for agricultural research for initiatives like AgMIP, CCAFS and Global Yield Gap Atlas, a vital contribution can made by releasing and visualizing relevant food security data to the broader research and policy community. From the past efforts, a number relevant aspects appear: 1. Use of standard data schema's and meta data where and when possible; 2. Invest time in user expectations to make sure the released data is taken up in work processes; 3. Ensure linking and consistency across domains and scales leading to an integrated image of food security data.

This paper describes our efforts in deriving a generic data schema for storing data on agricultural systems, covering several disciplines, temporal and spatial resolutions. Through this data schema we investigated the possibilities in storing and retrieving data for agricultural research in generic way across several project initiatives. As a case within the large domain of agricultural sciences, this paper uses data from crop experiments, or agricultural trials and inputs and outputs to cropping systems models, which are often derived or based on these crop experiments or agricultural trials. In our view, similar problems occur with other data sources in agricultural sciences, such as household data or sensor data, and potentially similar solutions in data schema and standards could be used.

The next section provides background on the domain of crop experiments and cropping systems models, projects or initiatives in which such models are used and an overview of some relevant theories from literature. The third Section describes the resulting data schema with some core innovations, which is followed by a short discussion and conclusion.

2 BACKGROUND AND METHODS

2.1 Crop experiments and cropping system models

Crop experiments or field trials are typically carried out at research stations or in some cases on farmers' fields, and the growing of the crop is closely followed by measuring a number of parameters (e.g. leaf area index, biomass, run-off, water content). Usually the crop is kept under controlled management, with which different variations of management options (e.g. no fertilizer vs. fertilizer, no irrigation vs. full irrigation with sprinklers vs drip irrigations) are investigated, leading to an insight of crop responses to changes in management, biophysical environment, pollution effects or breeding potential in plant breeding.

The data generated through such crop experiments can be used, among others, for parameterising and validating cropping system models, which are simulation models that simulate crop growth and development subject to management and environmental conditions for estimates of yield and environmental effects (Van Ittersum and Donatelli, 2003). Many different cropping system models exist, developed and used for different purposes across the globe. Typically these cropping systems models require input data on soil, climate, crop management and parameters describing the crop phenotype and genotype. Such models are typically used for studying responses to climate change, assessments of food security and effects of technology changes or improvement on crop productivity.
The use of crop experiment data in cropping systems models is hindered by the lack of easily available data in public repositories, structured according to a generally agreed upon schemata. This implies that individual researchers applying a cropping system model rely on their own network or institutional context, and have to investigate individually the meaning and significance of different parameters as recorded in the crop experiment through different methods, often obtained in diverse formats through colleagues.
Although it would be beneficial to have a mechanism for release of crop experiment data according to standards and agreed formats, there is a challenge in the semantic diversity of crop experiments and the biophysical conditions in which these occur. In crop experiments, many different management options (i.e. nutrient, water, pest, weed, conservation and tillage management can be studied, with many different intensities, and different measurement methods or recorded variables during the growing season. The biophysical environment in which such crop experiments occur can also be measured and described in many different ways, with especially for soils big differences leading difficulties in interpretation. For climate, usually the measurements and data are more homogenous, and easy to interpret.

2.2 Project context

In a number of different projects and initiatives the problem of management of crop experiment data for use in cropping system models is experienced. Instead of a dedicated solution per project, it should be possible to achieve synergies between the projects in compiling and releasing crop experiment data for further use in other projects. These projects are diverse in their purpose and aims:

1. The Agricultural Model Improvement and Intercomparison Project, www.agmip.org, is a distributed climate-scenario simulation exercise for historical model intercomparison and future climate change conditions with participation from multiple crop and agricultural economic modeling groups around the world. It targets improved assessments of global food security in relation to climate change for the IPCC future Assessment Reports (Rosenzweig et al., 2012). In this project, crop experiments are crucial for cropping systems model calibration, after which these can be used in large scale assessments of climate change impacts.
2. The Global Yield Gap Atlas is a project to compile an Atlas of yield gaps(i.e. difference between potentially possible yield and actually obtained yield), by simulating yield gaps at many different locations across the globe. Cropping system models are used to simulate yield gap. The project is bottom-up, meaning that local conditions have to be reflected as much as possible in the models, requiring well-documented crop experiments.
3. MOCCCASIN is a project focusing on monitoring of winter-wheat in Russia by improved modelling of winter-kill and satellite data assimilation. A module for winterkill is added to the WOFOST model (Boogaard, et al. 1998). A field dataset was compiled during an intensive field campaign in 2011. This data set has to be stored for the future for different types of analysis.
4. Joint Programming Initiative on Food, Agriculture, Climate Change and Environment is an agenda setting project of different EU member states, focusing on food security and climate change across crop science, economics and animal science. For crops, many different crop models are

planned to be compared through exercises of ensemble modelling, leaing to improvements and large scale assessments.

2.3 Methods

Crop experiments are diverse in set-up and environment, and their data should ideally be used in as many as possible cropping systems models, which are by themselves diverse in their configuration, theoretical approach and input data. This diversity of configurations, environments, measurements and models can be characterised as semantic heterogeneity (Bright, et al., 1994), which is a known challenge for database systems. Data (like models) in itself contains sophisticated statements of knowledge that ideally have to be opened up for scientists to use (Villa et al., 2009). Data modelling or data-driven modelling can be used to sketch relational diagrams explaining the semantic heterogeneity. As a more advanced representation, Villa et al (2009) propose ontologies (i.e. a specification of a conceptualization in concepts, relationships, properties and constraints (Gruber, 1993)), which are richer in their representation as entity relationship diagrams commonly used in data modelling.
In our research, a group of experts from the different projects was brought together to identify the common elements, and through several iterations developing a data model representing the semantic heterogeneity in crop experiments. The data model has been incorporated in an Microsoft Access database as a test, with small test data sets. The Microsoft Access database is a first prototype and in the next release, Microsoft Access will be dropped for more advanced solutions. As a future development, the data model will be converted into a relational database management system, and converted to an ontology to capture more of the knowledge and thinking.

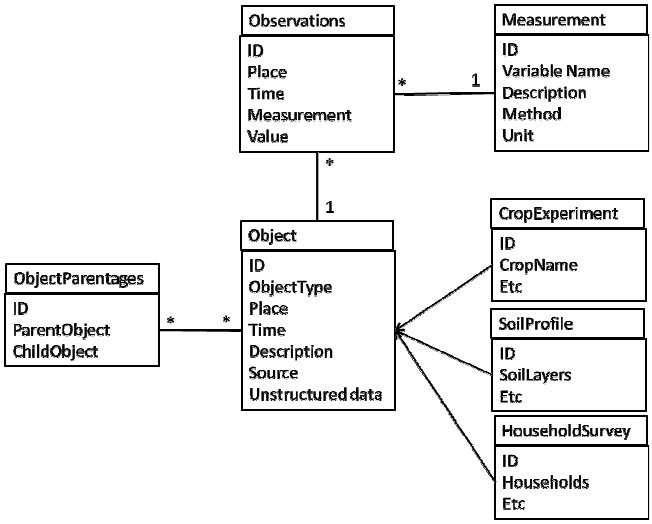

Figure 1. A simplified overview of the data schema for storing agricultural research data with central elements observations and objects.

3. RESULTS

3.1 Overview

The proposed data schema (Spatial Temporal Attribute Catalogue, STAC) is organised around an observation (Fig. 1), which is something being measured at a place and time through a method for measurement. This observation is combined with entities describing collections of observations as object, such a weather, crop, soil, farm, household, and with entities describing methods to measure observations as part of objects, such as an interview, weather station, soil profile,

pH meter, etc. The data schema as described here is not exhaustive, as more entities are required describing unique lists of relevant information, for example, crops, place, time, variables.
The data schema is loosely based on an evaluation of already existing data schema's in projects that could store the project specific information, and abstracts the common properties are presented in a more abstract level from the diversity of fields and set-ups in these projects, leading to the importance of the observation entity. This observation entity allows considerable flexibility in defining properties that are being observed at a time and space, allowing diversity of data. Also, an abstraction was made in defining space and time, allowing for inheritance of different types of space and time.

3.2 Observation Entity

This observation-entity was identified as the most generically possible element, as most data ultimately is derived from some sort of census, sensor, questionnaire or experiment, that occurs at a time and place, with a spatial and temporal validity. The observation-entity follows a key-value set-up, in which the definition of space, time and measurement are linked to values. The space, time and measurement definitions are drawn from other entities (Fig. 1) or drawn from standard referencing systems, such as OGC standards for space (not shown in Fig.1). In the measurement-entity of the data schema, variables (e.g. rainfall, pH, soil water content, farm income) are combined with methods to establish the value, the unit and a description. For example, rainfall could be with a unit of mm, with a description of annual rainfall in millimetres for a square meter, measured with a rain gauge.

3.3 Object entities

Observations can be combined to sets or groups in so-called objects (Fig. 1), that specify what is being measured on a higher aggregation level. For example, time series of observations on rainfall, radiation and wind speed combine to a description of weather for a location over a time period. The object would then be weather at location X for time period Y (Table 1).

Different types of such objects exists, which is handled through abstraction and inheritance, leading to different sub-objects inheriting common properties from the abstract object, allowing for flexibility in the definition of the child entities (Fig. 1). For example, an object describing a crop has different properties then an object describing weather. The crop object holds information on the cultivar and relevant parameters describing the phenotype or genotype, while the weather object holds information on the purpose of the type of weather station used.
Objects themselves can also be grouped through parent and child objects, for example a weather, soil, and crop object combine to a crop experiment parent object, given that they apply to the same location and time. In this way, more comprehensive data sets can be formed to describe complex real world data structures.

Table 1 A sample set up for the object entity, with meaning of the different columns part.

Field name (type)	Meaning
ObjectID (integer)	A unique number identifying the object, cannot be empty
ObjectTypeID (integer)	The type of the object (see later in this doc for explanation), cannot be empty
TimeStampBegin (timestamp)	The earliest value of the TimeStampBegin value of the ObjectDetails table for the object (
TimeStampEnd (timestamp)	The latest value of the TimeStampEnd value of the ObjectDetails table for the object or if missing the latest TimeStampBegin value of objectdetails
LongitudeDD (float)	The longitude of the object in decimal degrees, or the longitude of the point of gravitation of the GeoItem to which the data belong, cannot be empty
LatitudeDD (float)	The latitude of the object in decimal degrees, or the latitude of the point of gravitation of the GeoItem to which the data belong, cannot be empty
AltitudeM (float)	The altitude of the object in decimal degrees, or the altitude of the point of gravitation of the GeoItem to which the data belong, cannot be empty
GeoItemDef (yet unknown)	A external reference to a geoitem (point, line polygon), is allowed to be empty
Source (string)	A string describing the source of the object and its data, is allowed to be empty
Description (string)	A description of the object that is understandable by itself.
UnstructuredData (blob)	Anything relevant to further document the object and its data, is allowed to be empty

3.4 Technical Implementation

The STAC has been implemented in Microsoft Access™ as a first prototype, and methods were developed to load small test data sets in the STAC, to evaluate if the STAC can store the relevant data, as designed. In a technical implementation, it is foreseen that the STAC is largely 'hidden' to the outside world through the use of Application Programming Interfaces, to load, unload and visualize data (Fig. 2). Although the use of SQL is foreseen for data storage, the scripts themselves could become quite complex, requiring prepared views and stored proceduress that can be called through API's. An application has been developed to load and unload weather data from a structured file into the STAC, as a start and other applications will follow.

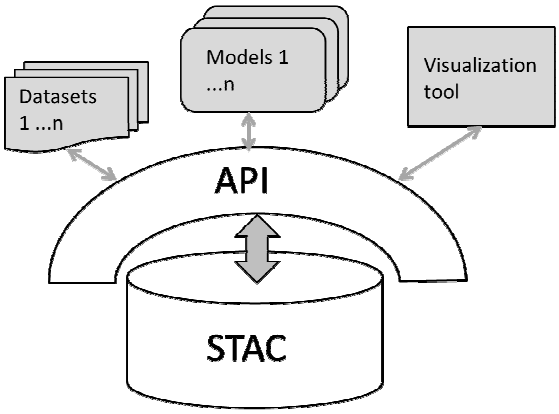

Figure 2. The foreseen architecture for the use of STAC

4 DISCUSSION AND CONCLUSION

The generic data schema proposed in this paper is an effort to coordinate or standardize the description and storage of data, at first instance for crop experiments, that should ultimately lead to an easier exchange of data among researchers and their research tools such as simulation models, statistical packages and visualization tools. Whether the proposed data schema will establish itself as a durable standard for the future, largely depends on its use in research projects, preferably with several partners involved. Two of the projects using the proposed schema are networking projects (Section 2.2), which involve more and more partners over time. This network development could stimulate the use of the proposed schema over institutes and researchers. Next to the use in projects, the link with existing standards will be explored, most notable the OGC/ISO Observations and Measurements (OM) standard (ISO/DIS 19156 www.opengeospatial.org/standards/om), to which the proposed data schema has resemblance in set up and philosophy.
A crucial step in its further development is therefore the use in projects, testing it, designing API's and supplying data to data management systems. The generic data schema will firstly be used in projects, largely focused on crop-related data, but other projects more focusing on households, soil and nutrient cycling, or biodiversity observations could also potentially use it, given the high level of abstraction in the data schema.
An expected challenge to the data schema will be the storage of large data sets of, for example, gridded climate data for large areas over a long time period. In such a case, the observation entity could end up containing billions of records, leading to slow performance to extract a specific subset of the data. Different solutions might be required for such data sets, such as distributed storage, or an extension of the currently foreseen data schema. The proposed data schema is designed for storing small but highly diverse data sets, such as crop experiments, and could therefore be most or only suitable to store that type of data, a priori excluding the large standardized datasets such as climate data.
A foreseen development of the STAC is to develop it into a semantic layer by employing an ontology, as proposed by Villa et al. (2009). With its key-value set up in some crucial entities, such as the observation entity, it might suit the storage in RDF of data according to a triple store, such as a Sesame database (www.openrdf.org). Having STAC as a semantic layer allows to save more information on the data schema, for example on relationships between entities and cardinality, and to release the STAC as an online ontology with URL's to which others can also link their developments.

ACKNOWLEDGMENTS

This work was funded through DFID-USDA funding for AgMIP, Bill and Melinda Gates Foundation funding for Global Yield Gap Atlas, EU DG-RTD for MOCCCASIN, and knowledge funds from the Dutch Ministry of Economics, Agriculture and Innovation.

REFERENCES

Berners Lee, T., Linked data-design issues, Retrieved on 2012 from *www.w3.org/DesignIssues/LinkedData.html,* 2012

Besemer, H., C. Addison, F. Pelloni, E. M. Porcari and N. Manning-Thomas, Agricultural Research, In: Verl, C. M. z. and W. Horstmann (Ed.) *Studies on Subject-Specific Requirements for Open Access Infrastructure*, Universitätsbibliothek, 19 - 68 pp., Bielefeld, 2011.

Bizer, C., T. Heath and T. Berners-Lee, Linked Data - The Story So Far, In: Heath, T., M. Hepp and C. Bizer (Ed.) *Special Issue on Linked Data, International Journal on Semantic Web and Information Systems (IJSWIS)*, 2012.

Boogaard, H. L., A.J.W. de Wit, J. te Roller and C.A. van Diepen, WOFOST 7.1. User's guide for WOFOST 7.1.3 crop growth simulation model and WOFOST control center 1.8., Technical Document 52, Wageningen, 2011. Available from http://www.wofost.wur.nl

Bright, M. W., A. R. Hurson and S. Pakzad, Automated resolution of semantic heterogeneity in multidatabases, *ACM Trans. Database Syst.*, 19(2), 212-253, 1994.

Hunt, L. A., G. Hoogenboom, J. W. Jones and J. W. White, ICASA Version 1.0 Data Standards for Agricultural Research and Decision Support, International Consortium for Agricultural System Applications Report, 37 pp., Honolulu, Hawaii, 2006.

Janssen, S., E. Andersen, I. N. Athanasiadis and M. K. Van Ittersum, A database for integrated assessment of European agricultural systems, *Environmental Science & Policy*, 12(5), 573-587, 2009.

Rosenzweig, C., J. W. Jones, J. L. Hatfield, A. C. Ruane, K. J. Boote, P. Thorburn, J. M. Antle, G. C. Nelson, C. Porter, S. Janssen, S. Asseng, B. Basso, F. Ewert, D. Wallach, G. Baigorria and J. M. Winter, The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and Pilot Studies, *Agricultural and Forest Meteorology*, In review, 2012.

van Evert, F. K., E. J. A. Spaans, S. D. Krieger, J. V. Carlis and J. M. Baker, A Database for Agroecological Research Data: I. Data Model, *Agronomy Journal*, 91(1), 54-62, 1999a.

van Evert, F. K., E. J. A. Spaans, S. D. Krieger, J. V. Carlis and J. M. Baker, A Database for Agroecological Research Data: II. A Relational Implementation, *Agronomy Journal*, 91(1), 62-71, 1999b.

Van Ittersum, M. K. and M. Donatelli, Special Issue of the European Journal of Agronomy: Modelling Cropping Systems, *European Journal of Agronomy*, 18(3-4), 187-394, 2003.

Villa, F., I. N. Athanasiadis and A. E. Rizzoli, Modelling with knowledge: A review of emerging semantic approaches to environmental modelling, *Environmental Modelling & Software*, 24(5), 577-587, 2009.

White, J. W. and F. K. van Evert, Publishing Agronomic Data, *Agronomy Journal*, 100(5), 1396-1400, 2008..