IJDC | *General Article*

Harmonizing the Metadata Among Diverse Climate Change Datasets

André Vellino
School of Information Studies
University of Ottawa

Abstract

One of the critical problems in the curation of research data is the harmonization of its internal metadata schemata. The value of harmonizing such data is well illustrated by the Berkeley Earth project, which successfully integrated into one metadata schema the raw climate datasets from a wide variety geographical sources and time periods (250 years). Doing this enabled climate scientists to calculate a more accurate estimate of the recent changes in Earth's average land surface temperatures and to ascertain the extent to which climate change is anthropogenic.

This paper surveys some of the approaches that have been taken to the integration of data schemata in general and examines some of the specific metadata features of the source surface temperature datasets that were harmonized by Berkeley Earth. The conclusion drawn from this analysis is that the original source data and the Berkeley Earth common format provides a promising training set on which to apply machine learning methods for replicating the human data integration process. This paper describes research in progress on a domain-independent approach to the metadata harmonization problem that could be applied to other fields of study and be incorporated into a data portal to enhance the discoverability and reuse of data from a broad range of data sources.

Received 15 January 2015 | *Accepted* 10 February 2015

Correspondence should be addressed to André Vellino, 55 Laurier Avenue East (11107), Ottawa ON, Canada K1N 6N5. Email: avellino@uottawa.ca

An earlier version of this paper was presented at 10th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

One of the critical features of a research data set is the metadata schema, sometimes referred to as the *data format*, that specifies the semantics for its data points. In fields of research such as demographics, a standardized metadata schema provides researchers the ability to integrate data that have diverse origins which, in turn, enables the analysis of time-series data and the capacity to extract, share and reuse data across disciplines. In addition, datasets whose metadata are standards-compliant are easier to discover and preserve. In short, a uniform metadata standard for datasets is of critical value not only to the researcher but also to the data curator.

However, datasets originating from research in differing disciplines may be related to one another but not available in a common format. For example, some data obtained by researchers in one discipline, such as ecology, may nevertheless be relevant to another discipline, such as climatology. Thus, a study in ecology whose data is organized for the purpose of understanding the correlation between climate change and the propensity of vegetation to fire could also be used as a component in a climate change simulation study.[1]

It is also possible that researchers within the same discipline have data that share common elements but are not all structured to comply with a common standard. It could be that researchers are insufficiently motivated comply with disciplinary standards; or it could be that the data originates from a variety of historical periods and geographic locations and that in each period and location the methods and practices for data collection were somewhat different.

Such is the case with the historical record of earth's climate. Many datasets were collected in different locations and at different periods in history, and each are organized and structured in different ways. Yet for statisticians and data researchers to discover historical trends, these data sets need to be integrated into a common data schema. Hence the heterogeneity of metadata schemata is a barrier to the proper analysis of climate datasets.

In "Climate data challenges in the 21st century", Overpeck et al. underscore the need for adequate metadata schemata to meet the needs of sharing and reuse of climate data:

> 'Although research scientists have been the main users of these [climate] data, an increasing number of resource managers (working in fields such as water, public lands, health, and marine resources) need and are seeking access to climate data to inform their decisions, just as a growing range of policy-makers rely on climate data to develop climate change strategies. Quite literally, climate data provide the backbone for billion-dollar decisions. With this gravity comes the responsibility to curate climate data and share it more freely, usefully, and readily than ever before' (Overpeck et al., 2011).

[1] As a concrete example, consider the study by Higuera, Briles and Whitlock (2014a) published in the Journal of Ecology and whose corresponding data was deposited in Dryad (Higuera et al., 2014b). The data was obtained from lake sediment cores containing pollen and charcoal that offer a history of vegetation and fired over a period of 6,000 years. The raw data files are interpreted by Matlab software to produce Excel spreadsheets whose columns are documented by a codebook text file.

 doi:10.2218/ijdc.v10i1.367

The objective of this paper is to describe an approach to the development of a tool that enables the automated harmonization of data schemata and fits into an automated data-integration process for managing and mechanizing the curation of research data.

The automated data-integration process envisioned in this paper is intended to become a component of the data management framework "Extract, Transform and Archive" (ETA), a model that was developed for managing and mechanizing the curation of research data (Vellino and Lemire, 2011). This model was itself inspired by the well-known "Extract, Transform and Load" (ETL) process model used in data warehousing to integrate heterogeneous data sources and enable uniform data analytics. Extract, Transform and Archive is an adaptation of ETL that addresses the specific needs of research datasets: discovery and archiving.

One important component of the ETA model is that it provides a framework for the transformation of data schemata from among heterogeneous data sources, particularly in situations where different user communities need to query the same data. The ETA data curation model distinguishes between four major classes of data transformation actions: (i) mergers and joins that enable the combination and integration of data from several sources; (ii) data cleaning, which identifies inconsistent data points; (iii) data filtering, whose function is to remove irrelevant data elements; and (iv) the task of aggregation and data mapping. While data that is integrated from different sources often performs at least one of those actions and possibly all of them, the focus of this paper is on the data integration element: the processes of merging diverse datasets by automatically generating metadata schemata that enable this merging.

A starting point for this study is the climate data that was integrated by the Berkeley Earth project.[2] This work acts as both an example of how such a metadata harmonization tool would need to function and as a gold standard for the training of a machine learning system that performs the integration.

The first part of this paper outlines some of the existing approaches to the data harmonization problem in the database and ontology communities. The second part examines some of the metadata characteristics of the earth-surface temperature source data integrated by Berkeley Earth, as well as the metadata characteristics of the Berkeley Earth destination data. It does so to describe the machine-learning approaches that apply to these datasets specifically, but that may also generalize to datasets in other fields.

Approaches to Data Harmonization

Several methods and strategies exist to address the problem of harmonizing diverse metadata formats. As indicated in the introduction, compliance to metadata standards is the simplest way to eliminate the problem. The development of crosswalks, such as those employed to enable the interoperability of bibliographic metadata standards, is perhaps the most frequently employed alternative. There are also automated and semi-automated methods, such database schema matching and ontology integration algorithms employed in business database warehousing and in semantic web research respectively. This section examines extant approaches to data harmonization and explains the motivation for a machine-learning approach.

2 Berkeley Earth: http://berkeleyearth.org/

Standards

The compliance-with-standards approach to the problem of metadata harmonization has been proposed by EarthCube[3] for geoscience data. This is also the approach for climate data taken by the NOAA's National Climatic Data Center[4]. Richard et al. (2014) describe one of EarthCube's objectives as "encouraging the development and adoption of community standards for Web interfaces to data, metadata and data formats, and software libraries."

Yet for climate data, there are many obstacles to compliance with metadata standards. One is the diversity of geographic and historical sources. For example, European and North-American data sources have significantly different ancestries and contemporary data collections contain a great deal more information (e.g. about vegetation and geography) than 18th century data.

Richard et al. express the difficulties of climate scientists with non-standard formats this way:

> 'Many legacy data issues stem from the difficulty that individual researchers, operating on limited budgets, experience in trying to curate data produced by their research. As a result, data documentation is commonly insufficient to enable cross-domain use or to repurpose data obtained from repositories. In addition, using nonstandard, heterogeneous data requires significant effort. The meaning of data may be unclear because of nonstandard vocabulary usage. Inconsistent practices for data sharing make each new data acquisition a time-consuming learning experience' (Richard et al., 2014).

In addition, climate datasets have a range of quality and reliability attributes and are also used for a variety of purposes, forcing climate scientists to issue dataset "products" that have been examined for quality and treated or data-cleaned to eliminate outliers and inconsistent data points. Notwithstanding current international efforts to establish global standards and unify data collections, the diversity of formats in climate datasets is still a present-day reality.

Crosswalks

The crosswalk method of mapping metadata schemas is best known for converting between bibliographic standards, such as MARC, Dublin Core and MODS (for a survey, see Haslhofer and Klas, 2010). For instance, a crosswalk mapping between MARC and MODS formats in XML could be implemented as an XSLT stylesheet that maps the elements of one schema into another.

The manual development of a crosswalk is a difficult and error-prone task that requires a detailed understanding of the associated metadata standards and the intended interpretation of their elements. While this approach is pragmatic and often used to obtain results in the short-term, it does not generalize well and the processes for developing crosswalks are neither automated nor easy to reproduce. One scaling issue with developing crosswalks is that if there are n related standards, it is necessary to develop $n(n-1)$ crosswalks to map each metadata standard into the other.

3 EarthCube: http://earthcube.org/
4 National Climatic Data Center: http://www.ncdc.noaa.gov/cdo-web/datatools

One way to automate the generation of crosswalks is to develop a domain-specific ontology-based architecture (Oldman and CRM Labs, 2014) for this domain. In this model, the ontology effectively acts as a "universal translation language" which reduces the crosswalk generation problem into two major mapping tasks – a mapping from one schema into the generic ontology and a mapping from the generic ontology into the other schema (Uschold and Gruninger, 1996). The complexity of the task of developing crosswalks for n schemata is then limited to developing only $2n$ mappings, one from each specific schema to the subject ontology and one from the subject ontology to each specific schema.

Given a suitably constructed ontology in a specific domain (e.g. the CIDOC Conceptual Reference Model, which provides a common semantic framework for cultural heritage information) it is possible to develop rule-based algorithms to generate candidate crosswalks between schemata (Gaitanou et al., 2012). However, this approach is only effective if the mediating translation schema is an adequate abstraction of the subject domain. Moreover, as the subject-domain evolves, the ontology changes and the corresponding rules in the algorithm need to be updated to generate the crosswalks.

In the domain of climatology the equivalent of crosswalks are often implemented directly in the software systems that perform the conversions from one format to another. For example, for the Berkeley Earth data integration effort there are 20 separate "data-loader" modules that ingest data stored in a variety of different formats and convert them into a common format.

Database Schema and Ontology Integration

The computer science community has had to deal with a similar problem: the problem of detecting the differences between and generating possible mappings across generic and heterogeneous data schemata. One need for these mappings arises in the context of business data integration and the other from the need to bring some semblance of order to the wild west of the semantic web. This problem is so acute for both communities that this topic has been an autonomous area of research since 2001 (Rahm and Bernstein, 2001).

In their survey of metadata integration methods, Bernstein, Madhavan and Rahm (2011) review several kinds of matching approaches and their implementations. For example, they discuss methods such as matching field names according to their linguistic similarity, content-based similarity methods that rely on the textual similarity of the fields' contents and methods that apply thesauri and dictionaries.

One important distinction among these various approaches is the one between schema-level and instance-level matching. Matching at the schema level uses information provided by the metadata schema, if they exist: schema elements are matched if they are similarly structured or have similar relationships to other similar elements. On the other hand, instance-level mapping uses information gleaned from the data contents, either in terms of their values or value-ranges and data types.

Ontology integration (Shvaiko and Euzenat, 2013) can be thought of as a special case of the schema-level matching problem applied to knowledge-based systems that express conceptual hierarchies using controlled vocabularies. Matching ontologies consists of performing alignments between entities in each ontology.

Existing tools for ontology integration are regularly assessed by the Ontology Alignment Evaluation Initiative[5] using test cases ranging from thesauri to biomedical

5 Ontology Alignment Evaluation Initiative: http://oaei.ontologymatching.org/

ontologies. Experimental results (Dragisic, Eckert, et al., 2014) show that fully automated matches have F_1-measures [$F_1 = 2$ (*precision* × *recall* / *precision* + *recall*)] that range from excellent (0.75) to poor (0.14), depending on the complexity of the matching problem. All of these systems are research efforts that are intended to apply to semantic web applications.

In summary, crosswalks, while a pragmatic solution for specific fields of endeavour, are hard to generalize and labour-intensive. Database schema matching systems and ontology integration systems typically rely on the known schemas source and target schemas to apply linguistic and rule-based approaches to perform the mapping. Neither of these strategies generalizes well, particularly in legacy data documentation environments whose interpretation is highly dependent on the software designed to read it, as is typically the case with climate datasets.

Earth Surface Temperature Datasets

For the layperson, the structure and organization of climate datasets is not obvious. For example, a typical data file from the NOAA's Global Historical Climatology Network (GHCN) repository[6] needs to be understood with reference to a code-book *readme* file that explains the encoding various data fields that exist in the different datasets (daily, monthly, blended).

The natural state of a NOAA's GHCN-D (daily) data file is as a plain text-file containing upwards of 270 characters per line, divided by convention into blocks by column number. Each line is a record of a climate-element reading at a location, such as a weather station: columns 1–11 designate the station ID, columns 12–15 contain the year, columns 16–17 the month and columns 18–21 the climate-element. Climate-elements consist primarily of temperature (max and min), precipitation and snowfall, but there are also dozens of other permissible element-codes for this column-block. Subsequent columns refer to measurement for the same variable on subsequent days for that month.

Static data, such as the location and attributes of weather stations, are found in separate files that are also organized in columnar format. Weather stations data have similar column conventions (e.g. columns 39–68 for the name of the weather station, columns 32–37 for the elevation, etc.) whose value-ranges and data-types (character, real, integer) are also specified in a separate metadata code-book file (Menne et al., 2012).

Thus a GHCN-D data file beginning with the characters "`USC00411646190408TMAX 356`" indicates that the maximum temperature in Channing, Texas in August 1904 was 35.6 Celsius. While it is quite simple to write software to parse these data into their component parts, it is another matter to teach a computer to recognize these regularities without explicit instructions.

GHCN-M (monthly) data contain monthly temperatures (min, max, avg) computed (where available) from daily minima and maxima. Included in a typical contemporary GHCN monthly dataset you can also find non-climate specific data fields that are often crucial to the interpretation of the climate-only data points (Lawrimore et al., 2011): the total population and population class in which the weather station is located centre (urban/suburban/rural); the topography surrounding the station (flat, mountainous etc.); the proximity of the station to a body of water; the distance of the station to an airport

6 GHCN Repository: http://www1.ncdc.noaa.gov/pub/data/ghcn/

 doi:10.2218/ijdc.v10i1.367

and up to 44 classes of proximate vegetation-types (bogs, shrubs, sand desert, tundra). Some of these fields also have functional dependencies on one another, such as total population and the population class-type.

Recent data integration efforts by the Global Land Surface Meteorological Databank (Rennie et al., 2014) now provide their data in alternative formats such as the NetCDF Climate and Forecast Metadata Convention (Eaton et al., 2014), which includes in the dataset files themselves the descriptors associated with each variable, such as temperature units and spatial coordinates as well as the default-values for null-data. Files in NetCDF-CF format also include provenance metadata.

Berkeley Earth Integration

The integration of earth surface temperature data sets by the Berkeley Earth project involves many steps that ultimately lead to a dataset collection that can be statistically analysed for different scientific purposes. But it is the first step (shown in Figure 1) with which the present paper is concerned: the transition from "Source Data – Different Formats" to the "Source Data – Common Format".

The common format chosen by Berkeley Earth strikes a pragmatic balance between simplicity of representation (space-separated columnar data in plain text prefaced by explanatory text in "%" comment format) and database normalization principles that factor out, for example, data about the weather stations and the temperature records for those stations. Information about the station data (weather station metadata, e.g. "Multiple names are associated with this site") varies according to the original sources and are recorded "flags" ("domains" in Database parlance).

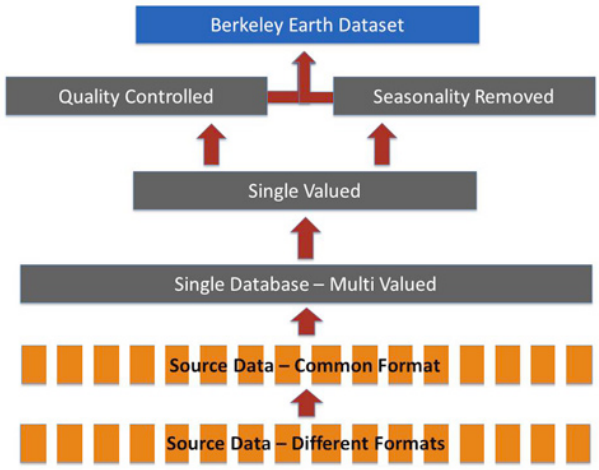

Figure 1. Berkeley Earth data integration flow chart.

All of the 20 different original data formats differ significantly from the Berkeley Earth destination format. As an example, consider some of the transformations that need to be learned by a machine learning system to convert GHCN-Daily (".dly") files to the Berkeley Earth common format. First, the initial 21 characters need to be parsed into GHCN Station ID | Year | Month | Element fields, as well as the 31 temperature value and flag column fields for each day of the month (step T1 in Figure 2). Secondly, each row that corresponds to a TMAX, TMIN and TOBS value in the Element field needs to be added to one of the three corresponding files (only two of which, TMAX and TMIN, are shown in step T2 in Figure 2). Third, there is the collapse of Year-Month (in GHCN

files) to a decimal-year representation in Berkeley Earth format (step T3 in Figure 2). In the Berkeley target format, the day on which a climate-data measurement was made is expressed as a year followed by a decimal fraction of a year that identifies the day using three digits of precision after the decimal point. In the source files, the day of the month is implicit in the column location of the temperature measurement and the month is explicit in the station-date-measurement field at the beginning of the record. Finally, there is the averaging of the Value1…Value31 (temperature) fields in the GHCN files that are reduced into a single "Temperature" field in the Berkeley Earth files (step T4 in Figure 2).

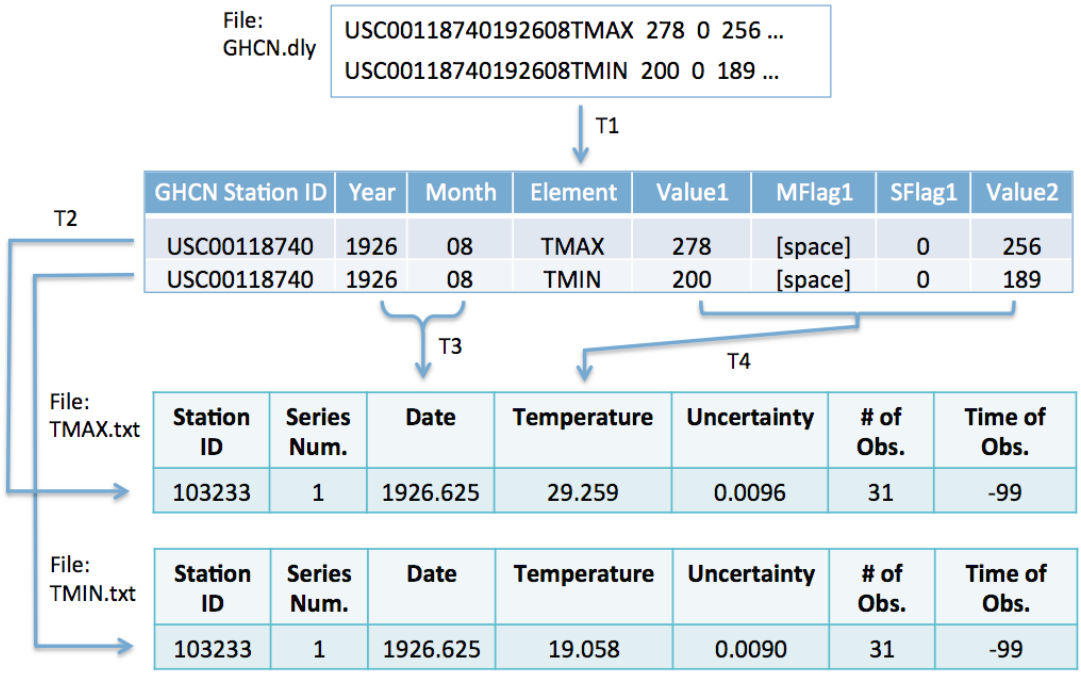

Figure 2. Some of the transformations from GHCN-Daily files to Berkeley Earth format.

While some of 20 different original formats format differ from one another only in subtle ways, other differences among them are significant. For instance, data files for CRUTEM3[7] (a text format that is now deprecated in favour of NetCDF) from the U.K. Met Office Hadley Centre, contain station information (Number, Name, Country, Latitude, Longitude, etc.) at the head of each file and the monthly temperatures in Celsius (with a decimal) for each year in 12 columns. The same data from GCOS Surface Network Monitoring Centre (GSNMC), on the other hand, obeys conventions for station identifier, temperature measurements (integers in deci-Celcius) and measurement-types similar to the ones in GHCN-M described above.

In short, the Berkeley Earth data harmonization was performed by a series of explicit crosswalks that were embedded in the data-conversion software. That multiple crosswalks have been written for the same target metadata schema affords the opportunity to automate the mapping method with machine learning methods.

7 CRUTEM3: http://www.metoffice.gov.uk/hadobs/crutem3/

 doi:10.2218/ijdc.v10i1.367

A Machine Learning Approach to Mapping Schemata

From among the Database Schema and Ontology Integration technologies mentioned above, the most promising for this task – and the one which has the potential to generalize to data-integration problems in other fields of data-intensive science – is supervised machine learning (Kotsiantis, 2007). It is well known that machine-learning algorithms are especially adept at solving classification problems, such as the assignment of class-labels to documents (e.g. 'spam'/'not-spam' labels to emails), and it is also evident that the metadata schema matching can be formulated as a classification problem (Doan, Domingos and Halevy, 2001). Thus algorithms like Support Vector Machines (SVMs) that have very high precision and recall (Abe, 2010) for document classification tasks hold a similar promise of performance for matching climate metadata.

The main challenge in demonstrating the practical value of this approach is to ensure that the training process effectively uses both the information in code-book (documentation) files and the regularities in data itself to avoid the need for human intervention (e.g. the Excel import wizard for files in comma separated values format). This requires breaking down the machine learning process into multiple stages that correspond to the manual transformations, such as those illustrated in Figure 2. The first task for the machine learning system is to deduce the boundaries of data fields that occur from data instances, e.g. to learn that character strings that appear to be compound terms (i.e. strings that do not have space or comma delimiters) should be decomposed into their significant components. To use the example above, expressions like "`USC00411646190408TMAX`" would need to be broken down into their constituent parts (Station ID | Year | Month | Measurement Variable) without explicitly writing code to do so. This kind of analysis could be done with a Conditional Random Field model for segmenting character sequences (Lafferty, McCallum, and Pereira, 2001).

The second step is to train the machine to recognize the implicit data-type transformations in the numeric data points, such as the combination of year-month (from the text in the row) and the day (from column position of the data point) in the source to the decimal representation of the date in the target format. In general there may be many such transformations that need to be learned (changes in coordinate systems, measurement units, etc.) It is not clear whether this can be done in a general way without providing the computer with rule-based heuristics about how to map numeric format representations to one another. However, in the worst case, explicitly enumerating a list of likely mappings for this task would be much easier and more reusable in other domains than segmenting compound character strings (the first step). The third step is to perform the label assignment by training, for example, an SVM to assign labels from destination data content fields in the destination codebook to source-data content fields that have been processed by steps one and two. Other tasks include learning the mapping of Station IDs, averaging multiple columns and identifying null-values.

Thus a combination of machine learning methods that can both data-mine the content and analyse the metadata files holds the promise of being an effective strategy for automating the data integration task in a manner that does not depend on any specific subject domain ontology.

Conclusions and Future Work

This paper argues for the value and the feasibility of a machine-learning approach for addressing the data harmonization problem. This approach makes it is easier to generalize to other data-organization paradigms and permits the mining of data files that do not conform to metadata schemata or ontologies expressed in a well-formed language (Database or XSD schema). A machine learning approach permits the exploitation of regularities in the data themselves to draw inferences using a collection of algorithms that are applied in a sequence.

There are some unknowns with this approach. It could be that some of the details of data encoding conventions, like the presence of "`-9999`" (or "`NA`" or "`N/A`" or even just a blank space) to signify the absence of a measurement, may not be easily detected as having this purpose without human intervention. Also, since this proposal has not yet been implemented or tested, it is not yet clear whether the relatively small (from a machine learning point of view) number of instances of mappings implemented by Berkeley Earth is sufficient to yield high enough precision results to enable an entirely automated process.

One important next step after a first implementation will be to test the machine learning system on similarly recorded (i.e. tabular) data from a different subject domain, such as ecology. Testing the validity of the machine model trained on climate data will require harvesting a collection of diverse datasets in ecology and a manual mapping of these datasets into a harmonized format.

Integrated into a discovery portal, such a system for data-mining, data and metadata in heterogeneous formats could support the discovery of new cross-disciplinary knowledge that is currently buried in data files whose organization is opaque to all but specialists in their field.

Acknowledgements

I wish to thank André Viau, Peter Turney, Daniel Lemire, Jim Elder and the anonymous reviewers for valuable comments on earlier drafts of this paper, as well Christine Newman for her research assistance.

References

Abe, S. (2010). *Support vector machines for pattern classification*. London: Springer. doi:10.1007/978-1-84996-098-4

Bernstein, P.A., Madhavan, J., & Rahm, E. (2011). Generic schema matching, ten years later. *Proceedings of the VLDB Endowment*, *4*(11), 695–701. Retrieved from http://www.vldb.org/pvldb/vol4/p695-bernstein_madhavan_rahm.pdf

Doan, A., Domingos, P., & Halevy, A.Y. (2001). Reconciling schemas of disparate data sources: A machine-learning approach. *ACM Sigmod Record 30*(2), 509–520. http://www.sigmod.org/sigmod/sigmod01/eproceedings/papers/Research-Doan-et-al.pdf

Dragisic, Z., Eckert, K., Euzenat, J., Faria, D., et al. (2014). *Results of the ontology alignment evaluation initiative 2014.* Retrieved from http://oaei.ontologymatching.org/2014/results/oaei2014.pdf

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Caron, J., … Graybeal, J. (2014). *NetCDF climate and forecast (CF) metadata conventions: Version 1.7.2 DRAFT, 28 March, 2014.* Retrieved from http://cfconventions.org/Data/cf-conventions/cf-conventions-1.7/build/cf-conventions.pdf

Gaitanou, P., Bountouri, L., & Gergatsoulis, M. (2012). Automatic generation of crosswalks through CIDOC CRM. In J. M. Dodero, M. Palomo-Duarte, & P. Karampiperis (Eds.), *Communications in Computer and Information Science: Volume 343. Metadata and Semantics Research* (pp. 264–275). doi:10.1007/978-3-642-35233-1_26

Haslhofer B., & Klas W. (2010). A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys 42*(2), Article 7. doi:10.1145/1667062.1667064

Higuera, P.E., Briles, C.E., Whitlock, C. (2014a). Fire-regime complacency and sensitivity to centennial- through millennial-scale climate change in Rocky Mountain subalpine forests, Colorado, U.S.A. *Journal of Ecology 102*(6), 1429–1441. doi:10.1111/1365-2745.12296

Higuera, P.E., Briles, C.E., Whitlock, C. (2014b). *Data from: Fire-regime complacency and sensitivity to centennial- through millennial-scale climate change in Rocky Mountain subalpine forests, Colorado, U.S.A.* [Data set]. doi:10.5061/dryad.q2b8t.2

Kotsiantis, S.B. (2007). Supervised learning: A review of classification techniques. *Informatica, 31*(3), 249–268. Retrieved from http://www.informatica.si/index.php/informatica/article/view/148

Lafferty, J., McCallum, A., & Pereira, F.C. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In C. E. Brodley & A. P. Danyluk (Eds.), *Proceedings of the 18th International Conference on Machine Learning (ICML '01)* (pp. 282–289). San Francisco, CA: Morgan Kaufmann.

Lawrimore, J.H., Menne, M.J., Gleason, B.E., Williams, C.N., Wuertz, D.B., Vose, R.S., & Rennie, J. (2011). An overview of the Global Historical Climatology Network monthly mean temperature data set. *Journal of Geophysical Research: Atmospheres, 116*(D19), Article 121. doi:10.1029/2011JD016187

Menne, M.J., Durre, I., Vose, R.S., Gleason, B.E., & Houston, T.G. (2012). An overview of the global historical climatology network-daily database. *Journal of Atmospheric and Oceanic Technology, 29*(7), 897–910. doi:10.1175/JTECH-D-11-00103.1

Oldman, D., & CRM Labs. (2014). *The CIDOC conceptual reference model (CIDOC-CRM) primer.* Retrieved from http://www.researchspace.org/file-cabinet/CRMPrimer_v1.1.pdf

Overpeck, J.T., Meehl, G.A., Bony, S., & Easterling, D.R. (2011). Climate data challenges in the 21st century. *Science, 331*(6018), 700–702. doi:10.1126/science.1197869

Rahm, E., & Bernstein, P.A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal*, *10*(4), 334-350. doi:10.1007/s007780100057

Rennie, J.J., Lawrimore, J.H., Gleason, B.E., Thorne, P.W., Morice, C.P., Menne, M.J., … Yin, X. (2014). The international surface temperature initiative global land surface databank: Monthly temperature data release description and methods. *Geoscience Data Journal, 1,* 75–102. doi:10.1002/gdj3.8

Richard, S.M., Pearthree, G., Aufdenkampe, A.K., Cutcher-Gershenfeld, J., Daniels, M., Gomez, B., Kinkade D., & Percivall, G. (2014). Community-developed geoscience cyberinfrastructure. *Eos, Transactions American Geophysical Union, 95*(20), 165–166. doi:10.1002/2014EO200001

Shvaiko, P., & Euzenat, J. (2013). Ontology matching: State of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, *25*(1), 158–176. doi:10.1109/TKDE.2011.253

Uschold, M., & Gruninger, M. (1996). Ontologies: Principles, methods and applications. *The Knowledge Engineering Review*, *11*(02), 93–136. doi:10.1017/S0269888900007797

Vellino, A., & Lemire, D. (2011). Extracting, transforming and archiving scientific data. In L. Candela, Y. Ioannidis, & P. Manghi (Eds.), *Proceedings of the Fourth Workshop on Very Large Digital Libraries (VLDL 2011)*. Retrieved from http://arxiv.org/abs/1108.4041