Supporting Tabular Data Characterization in a Large Scale Data Infrastructure by Lexical Matching Techniques

Leonardo Candela, Gianpaolo Coro, and Pasquale Pagano

Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo"
Consiglio Nazionale delle Ricerche
Via G. Moruzzi, 1 – 56124, Pisa, Italy
{candela,coro,pagano}@isti.cnr.it

Abstract. Digital Libraries continue to evolve towards research environments supporting access and management of multiform Information Objects spread across multiple data sources and organizational domains. This evolution has introduced the need to deal with Information Objects having traits different from those characterizing Digital Libraries at their early stages and to revise the services supporting their management. Tabular data represent a class of Information Objects that require to be efficiently managed because of their core role in many eScience scenarios. This paper discusses the tabular data characterization problem, i.e., the problem of identifying the reference dataset of any column of the dataset. In particular, the paper presents an approach based on lexical matching techniques to support users during the data curation phase by providing them with a ranked list of reference datasets suitable for a dataset column.

Keywords: tabular data management, data curation, large-scale data infrastructure, lexical similarity.

1 Introduction

Digital Libraries have evolved a lot during the last twenty years while maintaining and further strengthening their central role in knowledge sharing [6]. Digital Libraries are revolutionizing the whole knowledge management lifecycle. They are no longer perceived as a means to discover cultural heritage only, rather are nowadays conceived as innovative, dynamic, and ubiquitous research supporting environments. In such environments *communities of practice* [15,25] are expected to be able, through their Web browsers, to seamlessly access and exploit data, services, and processing resources managed by diverse systems in separate administration domains.

This evolution continues to enlarge the domains Digital Libraries are called to serve that presently include *eScience*, *cultural heritage*, and others [5,21,2,9,12,24]. Current Digital Library developers are called to develop complex systems that

M. Agosti et al. (Eds.): IRCDL 2012, CCIS 354, pp. 21–32, 2013.

have to give solutions to "traditional" issues, e.g., existing data providers federation, distributed retrieval, and long-term preservation, as well as "new" issues, e.g., social network models, large-scale computing, and micro information. Furthermore, they have to face scaled-up versions of the above issues with respect to various axes, e.g., number and variety of actors to be served, size and variety of content to be managed, and diversity of systems and technologies to be integrated. Very often the content they are requested to manage falls under the "*data*" category and their implementation actually requires the realization of Data Infrastructures.

The term "data" itself, although very common, is difficult to define since it may be given different meanings, both in the digital and in the real world. Actually, the act of recognising or understanding that "something" – e.g., observations, statistics, artefacts, records – constitutes data is an intellectual activity that is usually driven by a certain goal. Data is collected for many purposes, via different approaches and very often it is difficult to interpret once exploited in contexts other than its initial one [3,4]. Digital Libraries are called to manage data ranging from traditional research outputs, mainly papers and experimental data, to living reports [7,5], executable research papers [10,19], and scientific workflows [20]. Very often such data fall into the category of "*big data*" [23], i.e., data characterised by (*i*) *volume*, i.e., its dimension in terms of bytes is huge; (*ii*) *velocity*, i.e., its speed requirements for collecting, processing and using is demanding; and (*iii*) *variety*, i.e., its heterogeneity in terms of data types to be managed and data sources to be merged is high.

This paper discusses one of the problems arising when dealing with *tabular data*[1] management where management needs (*i*) to support collaboration among multiple users and organizations; (*ii*) to appeal to a broad audience of users who are not technically skilled; and (*iii*) to guarantee data completeness and correctness as to enable effective data analysis; i.e., to solve the problem of identifying, verifying and associating the actual controlled vocabularies that might have been used by the data provider while producing the dataset. Tabular data are mainly stored in CSV (Comma Separated Values) files where little or no emphasis is posed on representing and standardizing the characterization of the single columns they consist of. However, knowing the "type" of values a column is expected to contain (*controlled vocabulary*, *code list* or *reference dataset* rather than basic types such as string or integer) is a fundamental aspect when datasets have to be effectively managed for, e.g., certification of compliance, comparison, integration and analysis. To this aim, this paper proposes an approach for supporting an end user during the operations to transform a "*raw dataset*" – i.e., a dataset consisting of its data only – into a "*characterized dataset*" – i.e., a dataset where each column is characterized by the controlled vocabulary from which its values have been selected. The effectiveness of such an approach is discussed in the context of a Data Infrastructure.

[1] Tabular data is a very common format for a plethora of data ranging from observations to specimen records, catch statistics, surveys, etc.

The remainder of the paper is organized as follows. Section 2 characterizes the major challenges of the problem identified above. Section 3 describes the proposed approach. Section 4 assesses the effectiveness of the proposed approach. Finally, Section 5 concludes the paper and summarizes its results.

2 The Tabular Data Characterization Problem

Data-intensive science as well as approaches expecting to rely on data require three basic activities: data *capture*, *curation*, and *analysis*. In these scenarios, data come in all scales and shapes covering: large international experiments; cross-laboratory, single-laboratory, and individual observations; and also individuals lives [12].

In these settings it is fundamental to equip collected datasets with additional information aiming at characterizing each dataset and making it possible to interpret the dataset even in contexts other than its initial one. This additional information may range from *bibliographic*-oriented metadata to *provenance*-, *coverage*-, *certification*-oriented metadata. Enriched and standardized datasets are, in fact, simpler to be managed and allow for exploiting more predefined functionalities as to get high performances on analysis and processing.

Tabular data represent a very common format for many datasets in many different scenarios, e.g., statistical data, surveys, observations. A fundamental piece of information that should equip tabular data is the one characterizing the "*data type*" of any column a dataset contains. However, the actual notion of data type goes well beyond the expected ones like string or integer. In fact, the compilation of datasets commonly relies on existing *controlled vocabularies*, *code lists* and *reference datasets*[2]. For instance, in compiling a dataset on catch statistics or specimen records it is worth to refer to reference datasets for species names and zones. Such reference datasets usually contain a complete record for each of the instances the reference dataset is about, as well as links with other reference datasets. By linking a dataset with the reference datasets its values come from, the actual information contained in the dataset is multiplied. The motivations of this are similar to those of *Linked Data* [1].

Although reference datasets are used or alluded during datasets capture phase, any information about them is usually discarded when the tabular dataset is stored in a CSV file for management purposes. Moreover, this capture phase is usually performed in very diverse technological and organizational settings, thus leading to a very heterogeneous set of tabular datasets. Because of this, it is expected that a curation phase reconciles the "raw dataset" with its "characterized" / "curated" version when the datasets are aggregated in a common information space aiming at promoting their consumption.

Common issues that may arise when a user wants to "curate" a given dataset are the following:

[2] In the remainder of the paper the term reference dataset will be used to represent any dataset whose values are recognized instances of the elements the dataset is about, e.g., species, zones, countries.

- The raw dataset contains entries which might be misspelled with respect to the intended reference values;
- The raw dataset contains too many entries to be controlled by hand;
- The reference datasets are too many to be manually searched and then be associated with the dataset under curation;
- Potentially, many reference datasets might be associated to a given dataset (high level of ambiguity).

A complete comparison between a raw dataset and all the reference datasets would need high computational requirements. Moreover, it is not appropriate if a quick (almost real time) response time is expected, as it happens when the user is asking a web application to propose a reference dataset suitable for the dataset she/he is managing.

On the other hand, even a "*greedy*" approach is not so easy to identify because of the issues just discussed, e.g., a simple match between string data cannot be used because it is incapable to overcome the misspelling problems.

In the remainder of the paper, an approach for supporting an end user during the curation phase is proposed. It consists in an "helper" facilitating the identification of the reference datasets that have been actually used while compiling the "raw dataset". This approach is conceived to be fast and effective with respect to the issues discussed above.

3 An Approach for Tabular Data Characterization

The proposed approach is based on two algorithms: (*i*) a revised version of the Minimum Edit Distance (MED) and (*ii*) a constant complexity ranking procedure aiming at proposing a ranked list of suitable reference datasets given a column of a tabular dataset.

The Minimum Edit Distance (or Levenshtein Distance) algorithm was firstly introduced in [16]. It is a metric for measuring the amount of difference between two character sequences. It is defined as the minimum number of edits needed to transform one string into the other, the allowed edit operations being insertion, deletion, or substitution of a single character. The algorithm is based on a dynamic programming procedure introduced in [18] and has a computational complexity that is linear with respect to the product of the length (number of characters) of the strings to be compared. However, there exist several approaches for computing the "distance" between two strings or sequences of symbols. Some well known *similarity metrics*, i.e., measures for similarity or dissimilarity between two text strings for approximate matching or comparison, include: (*i*) the Hamming distance [11], which calculates the number of positions at which the corresponding symbols are different; (*ii*) the Needleman-Wunsch distance [18], which is used in bioinformatics to align protein or nucleotide sequences; and (*iii*) the Smith-Waterman distance [22], which is a variation of the previous one and performs local sequences alignment. Other techniques are used in various domains ranging from biology to phonetics, e.g., (*i*) the Jaro-Winkler distance [13,26], which is mainly used in the area of duplicates detection; (*ii*) the

Block or L1 distance [14], which introduces a new geometry for distance calculation, where Euclidean geometry is replaced by a new metric in which the distance between two points is the sum of the absolute differences of their coordinates; and (*iii*) the Soundex distance [17], which is a phonetic algorithm for indexing names by sound, as pronounced in English. Among the existing algorithms, we selected the MED one as it is the most common method for string comparison, its implementation is straightforward and it fits well with the characteristics of the proposed approach.

The constant complexity ranking procedure proposed is called *Lexical Guesser*. This is an approach defined by relying on the edit distance, which uses the lexical similarity scores for limiting the computational extent of the ranking procedure of a given column of a dataset. From that point on, a given column of a dataset which has been selected for curation purposes is called "*target dataset*". The Lexical Guesser uses similarities, instead of exact matching, in order to avoid to perform all the comparisons between the target dataset entries and all the entries of all the recognized reference datasets. The basic underlying ideas are:

- if the target dataset contains entries which are misspelled, errors can be recovered by using MED (actually, a revised version of it);
- if the target dataset is syntactically correct, then the computation can be limited by assuming that by picking some random chunks (samples) from the *right* reference dataset, these chunks will probably be lexically similar to the target dataset. For instance, a target dataset containing entries like 'North Atlantic Ocean', 'South Pacific Ocean', etc. would always get a non-minimal score when compared to the 'Oceans English Names' reference dataset because the latter also contains entries like 'Indian Ocean' or 'North Pacific Ocean' which share some lexical similarities with the target dataset entries. It is assumed that the recall of the search for the target dataset can include all those reference datasets presenting lexical similarities (over a certain *threshold*);
- the proposed approach is expected to be an helper for an activity that should remain semiautomatic, i.e., the algorithm reduces the search space of the possible reference data, while the final choice about the reference dataset to use is a duty of the user.

According to the above premises, the MED algorithm was modified and then incorporated into a ranking procedure realising the *Lexical Guesser*.

Actually, the MED algorithm has been enriched with a set of check rules and parameters aiming at enhancing its performances for the overall classification process. From a preliminary analysis, it has been noticed that the standard MED algorithm is not sufficient for calculating distances in the target scenarios. Some boosting rules have been added to raise or lower the scores in some cases.

The distance between two strings x and y is calculated as follows:

$$d(x,y) = \begin{cases} 0 & \text{if } \frac{maxln}{minln} > 1.5 \\ min\left(\frac{minln}{maxln} * 1.5, 0.9\right) & \text{if } contains(x,y) \vee contains(y,x) \\ 1 - \frac{MED(x,y)}{maxln} & \text{otherwise} \end{cases} \quad (1)$$

where:

- $maxln = max(length(x), length(y))$;
- $minln = min(length(x), length(y))$.

The constant values in the formula above are the result of an experimental activity. The limitation to 0.9 for the value of $d(x, y)$ when a string contains another one is a penalty score which aims to lower the distance value in the cases when strings are really close but not equal.

The ranking procedure consists in computing a similarity score $S(T, R_i)$ between the *target dataset* T, i.e., the values of a given dataset column, and every recognized reference dataset R_i as a product of three factors, namely (i) a *distance score* $D(T, R_i)$, (ii) a *coverage score* $C(T, R_i)$, and (iii) a *weight score* $W(R_i)$, by actually relying on samples of both the datasets, i.e., $\overline{T}$ and $\overline{R_i}$. A score α is computed to estimate the representativeness of the sample $\overline{R_i}$ as follows: $\alpha = |\overline{R_i}|/|R_i|$.

Given a *target dataset* T, for each reference dataset R_i the similarity score $S(T, R_i)$ is calculated by the following formula:

$$S(T, R_i) = D(T, R_i) * C(T, R_i) * W(R_i) \tag{2}$$

where

1. the *distance score* $D(T, R_i)$ is computed as the average distance between all the pairs of the selected samples $\{(t_k, r_{i_j})|t_k \in \overline{T} \wedge r_{i_j} \in \overline{R_i}\}$ where the distance is greater than an "acceptance treshold" τ as follows:

$$D(T, R_i) = \frac{\sum \{d(t_k, r_{i_j})|d(t_k, r_{i_j}) > \tau\}}{|\{(t_k, r_{i_j})|d(t_k, r_{i_j}) > \tau\}|} \tag{3}$$

2. the *coverage score* $C(T, R_i)$ is computed by multiplying the α score aiming at indicating the representativeness of the sample $\overline{R_i}$ by a factor aiming at indicating the similarity between $\overline{R_i}$ and $\overline{T}$ as follows:

$$C(T, R_i) = \alpha * \frac{S}{|\overline{R_i}|} = \frac{S}{|R_i|} \tag{4}$$

   where $S = |\{r_{i_j}|r_{i_j} \in \overline{R_i} \wedge \exists t_k|t_k \in \overline{T} \wedge d(r_{i_j}, t_k) > \tau\}|$
3. the *weight score* $W(R_i)$ is computed (i) by comparing the "size" of R_i with respect to the size of the whole set of recognized datasets and (ii) mitigating the impact of "big" dataset via logarithmic transformation as follows:

$$W(R_i) = \begin{cases} \frac{|R_i|}{\sum |R_j|} * 100 & \text{if } \frac{|R_i|}{\sum |R_j|} * 100 \leq 1 \\ \log\left(\frac{|R_i|}{\sum |R_j|} * 100\right) & \text{otherwise} \end{cases} \tag{5}$$

It is evident that the higher is each factor value, the higher the similarity score. This means that a very high score could imply a good overall similarity among

the single entries but even that the elements in T cover a big percentage of the R_i set.

Given a *target dataset* T, the list of recommended reference datasets is produced by sorting the set of reference dataset according to the values of the similarity score $S(T, R_i)$ and pruning those whose score differs from the top-ranked element in the list (i.e., the best score) for more than a given customizable threshold (Maximum Difference from Best Threshold or MDBT).

Overall, the complexity of the procedure depends from the number of string comparisons to be performed. If k is the number of reference datasets recognized and $|\overline{T}| = m$ and $\forall i, |\overline{R_i}| = n$, then the overall number of comparisons to be performed is $k * m * n$. However, because of its characteristics, the proposed approach is incline for parallelization both with respect to the reference datasets (every $S(T, R_i)$ can be calculated by an independent process) as well as with respect to the single reference dataset (independent processes can be used to calculate factors of the same $S(T, R_i)$).

The procedure can then be tuned in order to get results in an acceptable time, e.g., by establishing proper values for n and m as well as for the thresholds and the rest of parameters discussed above. The higher is the number of comparisons, the higher will be the complexity of the calculation as well as the accuracy. These aspects are discussed in the next Section.

4 Experiment and Results

The experiment we performed to validate the approach is based on tabular datasets and reference datasets expected to be managed in the context of the large scale data infrastructure implemented by D4Science and D4Science-II projects [8]. In particular, the settings are those resulting from an environment aiming at providing fisheries statisticians with a set of tools to manage tabular data on catch statistics. Tabular data usually are time series coming from observations about fishery periodic catches in terms of quantities and costs. When an user wants to manage a new time series, in order to use all the facilities offered by the environments for time series analysis and consumption, she/he has to curate such dataset by recognizing the reference datasets it exploits. Such operation involves the correction of misspelled entries, the identification of the data types for the columns and a validation of the coherence of the dataset contents. In this phase, the user is expected to rely on facilities helping the identification of the most suitable reference datasets. These facilities are based on the Lexical Guesser.

In this scenario, the set of recognized reference datasets is about information on marine species, e.g., species names, geographical areas, economic zones. It consists in 326 reference datasets, containing from 5 to 39,000 elements. These reference datasets can be classified as follows:

- *no overlap* – reference datasets that are *disjoint* from each other;
- *medium overlap* – reference datasets that present a *medium degree of intersection* with other ones, i.e., 20-50% of their entries overlap with entries

in at least another reference dataset (e.g., FAO area names, the ocean and sub-ocean names and the geographical names);
- *high overlap* – reference datasets that have a *large degree of intersection* with others, i.e., 80-90% of their entries overlap with entries in at least another reference dataset (e.g., species names coming from different species databases).

Each experiment reported in the remainder of this paper was executed by using 50 different target datasets for 20 times per input. The average score is reported in the tables.

In order to test the performances of the proposed approach, the following well known measures have been exploited:

$$Accuracy = \frac{TrueNegatives + TruePositives}{TotalNumberofReferenceDatasets} \tag{6}$$

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{7}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{8}$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{9}$$

where:

- *True Negatives* indicates the number of classifications which are correctly classified as *not suitable*,
- *True Positives* indicates the number of reported classifications which are really suitable for labeling an unknown target dataset;
- *False Negatives* and *False Positives* are defined by complement of the above.

The experiment configuration was set as follows:

- one sample of 25 elements was taken from a target dataset;
- a sample of 625 elements was taken from each reference datasets;
- the threshold for pruning the ranked list of $S(T, R_i)$, i.e., MDBT, was set to 30%;

The performance of the following tree approaches have been assessed:

- *Lexical Guesser* – i.e., the approach proposed in Sec. 3, based on the ranking of similarity scores $S(T, R_i)$ by formula 2;
- *Simple Matcher - Constant Complexity* – i.e., an approach based on the same ranking procedure (with pruning) where the distance $d(x, y)$ is based on exact string matching;
- *Simple Matcher - High Complexity* – i.e., an approach based on the ranking of $S(T, R_i)$ for all the reference datasets where the distance $d(x, y)$ is based on exact string matching.

Table 1. Results on a column of a dataset that exactly matches a reference dataset (results are expressed in percentages)

	Lexical Guesser	Simple Matcher Constant Complexity	Simple Matcher High Complexity
No overlap			
Accuracy	100	100	100
Precision	100	100	100
Recall	100	100	100
F1	100	100	100
Medium overlap (20%-50%)			
Accuracy	99.18	99.18	99.40
Precision	28.89	28.89	38.89
Recall	100	100	100
F1	44.44	44.44	44.44
High overlap (80%-90%)			
Accuracy	99.77	99.85	99.92
Precision	70.83	75	87.50
Recall	100	100	100
F1	79.17	83.33	91.67

Table 1 reports the results for target datasets that match exactly one reference dataset. In the case of *no overlap*, all the approaches get a 100% of accuracy. The task is quite trivial and the ranking procedure with pruning does not influence the performances. In the case of *medium overlap*, the 'Simple Matcher - High Complexity' approach is expected to perform better than the others, while errors are experienced with the 'Simple Matcher - Constant Complexity'. The 'Lexical Guesser' performs as good as the 'Simple Matcher - High Complexity', thus the flexibility of $d(x, y)$ does not help in this case. In the case of *medium overlap*, the Lexical Guesser performs worst than the others, however the performances are still acceptable for the application scopes.

Table 2 presents the performances when the experiment focuses on target datasets containing misspelled entries and entries that do not occur at all in reference datasets for the 50 to 100% of their entries. The 'Lexical Guesser' always outperforms the other two approaches. Moreover, the 'Simple Matcher - Constant Complexity' introduces errors. Performances are appreciable both in terms of accuracy and recall, which means that the approach is always able to return the right reference datasets to the user. The recall of approaches based on simple matching is always lower than that of the Lexical Guesser because in some cases the target dataset may be ambiguous, so that more than one reference dataset is suitable for it. In this case the choice necessarily is on the user's side, as she/he only knows the real nature of her/his data. A simple match tends to find few columns, while the proposed approach uses the flexibility of the comparisons in order to propose more reference datasets. The precision score indicates that in presence of either low or high ambiguity, the Lexical Guesser is able to extract the correct information, while with medium ambiguity, the

Table 2. Results on a column of a dataset which does not match exactly reference datasets (results are expressed in percentages)

	Lexical Guesser	Simple Matcher Constant Complexity	Simple Matcher High Complexity
No Superpositions			
Accuracy	100	99.69	94.89
Precision	100	66.67	35.29
Recall	100	44.44	55.56
F1	100	53.33	30.37
Medium Superpositions (20%-50%)			
Accuracy	99.54	99.39	99.54
Precision	58.33	100	100
Recall	100	45.83	62.50
F1	73.33	60	70
High Superpositions (80%-90%)			
Accuracy	99.54	99.23	99.23
Precision	100	50	50
Recall	50	16.67	16.67
F1	65	25	25

statistical nature of the algorithm begins to be evident. This happens because for some samples lexical similarities are found, while for others they are not retrieved. As for the F1 measure, it can be noted that it gives an estimation of the overall functioning, and the value for the Lexical Matcher is always higher.

5 Conclusion

The evolution of Digital Libraries calls for innovative, dynamic, and ubiquitous research supporting environments where communities of practice can seamlessly access data, software, and processing resources managed by diverse systems in separate administration domains through their Web browsers. In these environments data are multiform and their management demand for new methods.

This paper has discussed one of the problems arising when dealing with tabular data management where management occurs in scenarios characterized by these needs: (*i*) supporting collaboration among multiple users and organizations; (*ii*) appealing to a broad audience of users who are not technically skilled; and (*iii*) guaranteeing data completeness and correctness as to enable effective data analysis; i.e., giving solution to the problem of identifying, verifying and associating the actual reference datasets that might have been used by the data provider while producing the dataset.

It has been proposed an approach supporting an end user during the massaging of a "*raw dataset*" to transform it into a "*characterized dataset*" defined by associating the proper reference datasets that might have been used while capturing the data. This approach is based on (*i*) a similarity measure aiming

at estimating the similarity among the entries of the target dataset and the entries of the reference dataset by overcoming misspelling issues and (*ii*) a ranking approach appropriate for a real time use and aiming at providing the end user with a sorted list of reference datasets suitable for a given target dataset.

The experimental results show that the proposed approach actually outperforms other approaches in presence of misspelled entries, even if it looses in performances with respect to an approach based on exact string matching when user's data completely agree with some of the reference dataset.

Acknowledgments. The work reported has been partially supported by the *D4Science-II* project (FP7 of the European Commission, INFRA-2008-1.2.2, Contract No. 239019) and the *iMarine* project (FP7 of the European Commission, FP7-INFRASTRUCTURES-2011-2, Contract No. 283644). The authors would like to thank M. B. Baldacci (ISTI-CNR) for many helpful comments.

References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. International Journal on Semantic Web & Information Systems 5(3), 1–22 (2009)
2. Blanke, T., Candela, L., Hedges, M., Priddy, M., Simeoni, F.: Deploying general-purpose virtual research environments for humanities research. Philosophical Transactions of the Royal Society A 368, 3813–3828 (2010)
3. Borgman, C.: Research data: Who will share what, with whom, when, and why? In: China-North America Library Conference, Beijing (2010)
4. Borgman, C.: The Conundrum of Sharing Research Data. Journal of the American Society for Information Science and Technology, 1–40 (2011)
5. Candela, L., Akal, F., Avancini, H., Castelli, D., Fusco, L., Guidetti, V., Langguth, C., Manzi, A., Pagano, P., Schuldt, H., Simi, M., Springmann, M., Voicu, L.: DILIGENT: integrating Digital Library and Grid Technologies for a new Earth Observation Research Infrastructure. International Journal on Digital Libraries 7(1-2), 59–80 (2007)
6. Candela, L., Castelli, D., Pagano, P.: History, Evolution and Impact of Digital Libraries. In: Iglezakis, I., Synodinou, T.-E., Kapidakis, S. (eds.) E-Publishing and Digital Libraries: Legal and Organizational Issues, ch. 1, pp. 1–30. IGI Global (2011)
7. Candela, L., Castelli, D., Pagano, P., Simi, M.: From Heterogeneous Information Spaces to Virtual Documents. In: Fox, E.A., Neuhold, E.J., Premsmit, P., Wuwongse, V. (eds.) ICADL 2005. LNCS, vol. 3815, pp. 11–22. Springer, Heidelberg (2005)
8. Castelli, D.: D4Science-II - An e-Infrastructure Ecosystem for Science. ERCIM News 79, 9 (2009)
9. Crane, G., Babeu, A., Bamman, D.: eScience and the humanities. International Journal on Digital Libraries 7(1-2), 117–122 (2007)
10. Gorp, P.V., Mazanek, S.: SHARE: a web portal for creating and sharing executable research papers. Procedia CS 4, 589–597 (2011)
11. Hamming, R.W.: Error detecting and error correcting codes. Bell System Technical Journal 29(2), 147–160 (1950)

12. Hey, T., Tansley, S., Tolle, K.: The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research (2009)
13. Jaro, M.A.: Advances in record linkage methodology as applied to the 1985 census of tampa florida. Journal of the American Statistical Society 84(406), 414–420 (1989)
14. Krause, E.F.: Taxicab Geometry. Dover Publications (1987)
15. Lave, J., Wenger: Situated Learning: Legitimate Peripheral Participation. Cam (1991)
16. Levenshtein, V.: Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady 10, 707–710 (1966)
17. National Archives and Records Administration. The Soundex Indexing System (2007)
18. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48(3), 443–453 (1970)
19. Nowakowski, P., Ciepiela, E., Harezlak, D., Kocot, J., Kasztelnik, M., Bartynski, T., Meizner, J., Dyk, G., Malawski, M.: The collage authoring environment. Procedia CS 4, 608–617 (2011)
20. Roure, D.D., Goble, C.A., Stevens, R.: The design and realisation of the $\text{my}_{\text{experiment}}$ virtual research environment for social sharing of workflows. Future Generation Comp. Syst. 25(5), 561–567 (2009)
21. Shen, R., Vemuri, N.S., Fan, W., Fox, E.A.: Integration of complex archaeology digital libraries: An ETANA-DL experience. Information Systems 33(7-8), 699–723 (2008)
22. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. Journal of Molecular Biology 147, 195–197 (1981)
23. Stapleton, L.K.: Taming Big Data. IBM Data Management Magazine 16(2), 12–18 (2011)
24. Wallis, J.C., Mayernik, M.S., Borgman, C.L., Pepe, A.: Digital libraries for scientific data discovery and reuse: from vision to practical reality. In: Proceedings of the 10th Annual Joint Conference on Digital Libraries, JCDL 2010, pp. 333–340. ACM, New York (2010)
25. Wenger, E.: Communities of Practice: Learning, Meaning and Identity. Cambridge University Press (1998)
26. Winkler, W.E.: String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In: Proceedings of the Section on Survey Research Methods (American Statistical Association), pp. 354–359 (1990)