Summary-based Comparison of Data Quality across Public MAGE-ML Genomic Datasets

Lorena Etcheverry[1], Mariano P. Consens[2]

[1] Instituto de Computación, Facultad de Ingeniería, Universidad de la República
lorenae@fing.edu.uy
[2] University of Toronto
consens@cs.toronto.edu

Abstract. Extensive microarray experimental data is available online, facilitating independent evaluation of experiment conclusions and enabling reuse. Numerous microarray experiment datasets are published using the MAGE-ML XML schema but assessing the quality of published experiments still represents a challenging task since there is no consensus among microarray users on a framework to measure datasets quality. In this article, we apply techniques based on DescribeX that quantitatively and qualitatively analyze MAGE-ML public collections, gaining insights about schema evolution. Our case study shows that DescribeX is a useful tool for the evaluation of microarray experiment data quality that enhances the understanding of the instance-level structure of MAGE-ML datasets and its evolution.

Categories and Subject Descriptors: H. Information Systems [**H.m. Miscellaneous**]: Databases

Keywords: XML, data quality

1. INTRODUCTION

The collection of genes that are expressed or transcribed from genomic DNA, called transcriptome, is a major determinant of cellular phenotype and function. In the context of human health and treatment, gene expression measurement can help determine the causes and consequences of disease and how drugs work in cells and organisms [Lockhart and Winzeler 2000]. DNAmicroarrays are devices that measure the expression of genes, leading to a qualitative change in the ability to understand regulatory processes occurring at the cellular level that has revolutionized molecular biology and medicine [Kohane et al. 2002]. Gene expression experiments with microarrays are complex processes with multiple sources of variability. The functional genomics and informatics community had come together in order to develop common data exchange formats for gene expression experiments. The Microarray Gene Expression Database Society (MGED) put forward a proposal in 2004 for an XML-based format for exchanging this experiments, called Microarray Gene Expression - Markup Language (MAGE-ML) [Spellman 2002]. Although simpler exchange formats have been proposed[1] [Rayner 2009], MAGE-ML is still used and several gigabytes of experiments are publicly available in this format in experiments repositories such as ArrayExpress [Brazma 2003], caArray [Bian 2009], CIBEX [Ikeo et al. 2003] and SMD [Demeter 2007].

While XML provides flexibility for data providers to define their own attributes, it is also responsible for heterogeneity in data from different research groups. Differences in schema usage patterns, for example the use of optional parts of the schema, may lead to better quality levels in certain data

[1] MINiML, MIAME Notation in Markup Language http://www.ncbi.nlm.nih.gov/geo/info/MINiML.html

sources. Visual exploration of the schema usage may be very helpful, since it reveals the actual structure of a collection at the instance level, element usage frequency, consistency and general patterns of usage. Understanding how the schema usage evolves helps the scientific community to answer questions which are not possible by knowledge of the schema alone or by data browsing. Even though public databases use standard exchange formats [Do et al. 2003], most of them do not asses the quality of the published data [Brazma 2003]. Incomplete or unprecise experimental descriptions endangers the reuse of published experiments. The existence of quality experimental metadata is crucial for scientifics to decide on the reproducibility and veracity of published results [Brettschneider et al. 2008].

Data quality assessment of gene expression experiments is a challenging task since the community of microarray users has not yet agreed on a framework to measure quality in microarray experiments. User expectations with respect to the level of gene expression data quality vary substantially. Quality levels can depend on time frame and financial constraints associated with experimental effort, as well as the purpose of the data collection [Brettschneider et al. 2008].

The functional genomics and informatics community has made extensive microarray experimental data available online, facilitating independent evaluation of experiment conclusions and enabling reuse. Numerous microarray experiment datasets are published using the MAGE-ML XML schema and assessing the quality of published experiments is a challenging task since there is no consensus among microarray users on a framework to measure datasets quality [Allison et al. 2006].

In this article, we apply techniques based on DescribeX that quantitatively and qualitatively analyze MAGE-ML public collections, gaining insights about schema evolution through time and comparing data quality within different data sources [Etcheverry et al. 2010].

Our work shows that visualization techniques and quantitative schema usage analysis supported by DescribeX can be used to explore data quality in a collection. In particular, we generate insights into MAGE-ML standard usage evolution and quality of the published experiment collection. This work offers a new approach and tool to the microarray community for managing, monitoring, and growing the MAGE-ML standard.

The rest of the article is organized as follows. In Section 2, we review existing literature in data quality assessment of semi-structured biological data. In Section 3, we present key aspects in MAGE-ML data collections, and how DescribeX can be used to explore its quality. Later, in Section 4, we show how DescribeX can be used to analyze schema usage evolution. In Section 5, we present the results of measuring controlled vocabulary usage and the quality of the references, followed by the conclusion in Section 6.

2. RELATED WORK

There are several works on biological and biomedical data quality assessment. In [Müller and Naumann 2003], production of genome data is analyzed and several error types are categorized. In [Martinez and Hammer 2005], data quality measures are defined over sequence databases such as GeneBank and EMBL. The Qurator project [Missier et al. 2007] proposes a generic data quality framework based on users quality views, and its application to microarray experiments. However, the Qurator project proposal focuses on capturing user notion of quality rather than proposing microarray experiments quality metrics.

Data quality research has mostly focused on structured-data quality. Techniques for managing and improving data quality in semi-structured and unstructured formats are needed, as stated in [Madnick et al. 2009]. In [Goldman and Widom 1997] dataguides were introduced to help understand the structure present in semi-structured data collections.

Visual exploration could be used to gain insights from large scientific datasources [Gray et al. 2005].

These kinds of analysis in large data collections are non trivial and can be helped by summarization, which is not commonly supported by conventional XML tools. In particular, DescribeX [Consens et al. 2008; Ali et al. 2008] has been used to explore schema usage in protein-protein interaction XML datasets [Samavi et al. 2007].

In this article we extend the work presented in [Etcheverry et al. 2010], showing that DescribeX can be used to explore MAGE-ML standard usage evolution and quality of the published experiment collection. This work offers a new approach and tool to the microarray community for managing, monitoring, and evolving the MAGE-ML standard.

3. DATA SOURCES AND THE DESCRIBEX TOOL

In Section 3.1 we provide an overview of the MAGE-ML data sources explored and in Section 3.2 we provide an overview of DescribeX.

3.1 Data Sources

Even though several public repositories of gene expression experimental data exist, ArrayExpress [Brazma 2003] and Gene Expression Omnibus (GEO) [Edgar et al. 2002] are the most popular. Figure 1a shows the size of repositories, measured in quantity of published experiments and taking into account all species. Since 2009 NCBI has been importing experimental data from GEO into ArrayExpress. This allows users to view and search GEO and ArrayExpress data from a common interface and access the data in the standard MAGE-ML format [Parkinson 2009]. Figure 1b displays ArrayExpress evolution in terms of available experiments and size over the last seven years, showing that GEO-imported experiments have significantly increased the amount of experiments published in ArrayExpress [2].

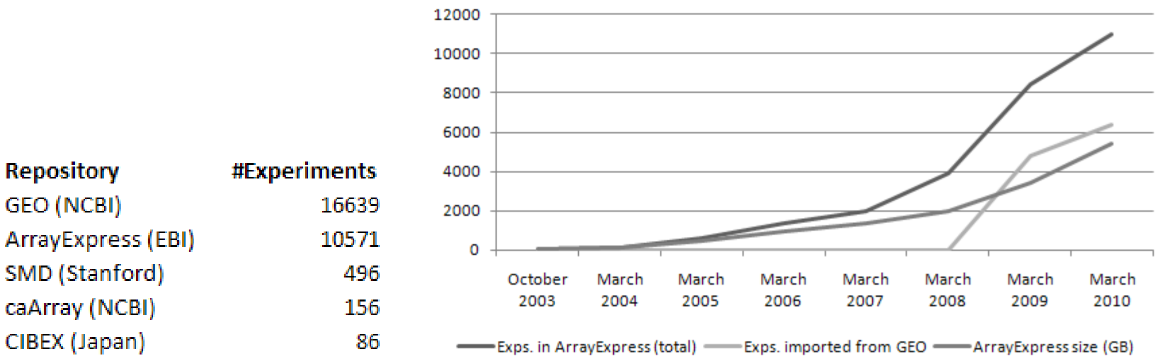
In this work, the data quality of more than 440 different datasets from the ArrayExpress repository are examined. Figure 2 shows the proportion that represent selected experiments for each data source and year of publication, where GEOD experiments are experiments imported from GEO and MEXP experiments are experiments originally submitted into ArrayExpress. The selected datasets are provided by different laboratories that examine *Homo Sapiens* using Affymetrix *HG-U133A* microarrays and jointly represent almost 20% of the publicly available experiments stored in ArrayExpress that use that particular chip over *Homo Sapiens* [3].

3.2 DescribeX

DescribeX is an application to construct and explore XML structural summaries [Ali et al. 2007; Consens et al. 2008]. A structural summary is constructed by grouping elements that have equivalent neighborhoods. The neighborhood is specified using an Axis Path Regular Expression (AxPRE), which is a connected subgraph originating from each XML element in the XML graph obtained by traversing XPath axes. Each P* summary groups elements that have the same label path to the root, i.e., obtained by traversing the (P)*arent* axes from the element to the root [Etcheverry et al. 2010]. A DescribeX structural summary is visualized as a graph with labeled edges and labeled nodes. DescribeX also implements *Coverage*, which represents a way to show the most relevant structures in an instance collection by hiding non-relevant portions from its summary. Coverage can be applied at a range of values from 0% to 100%.

[2] Data obtained from http://www.ebi.ac.uk/microarray/doc/stats/

[3] Last accession to ArrayExpress on April 2010

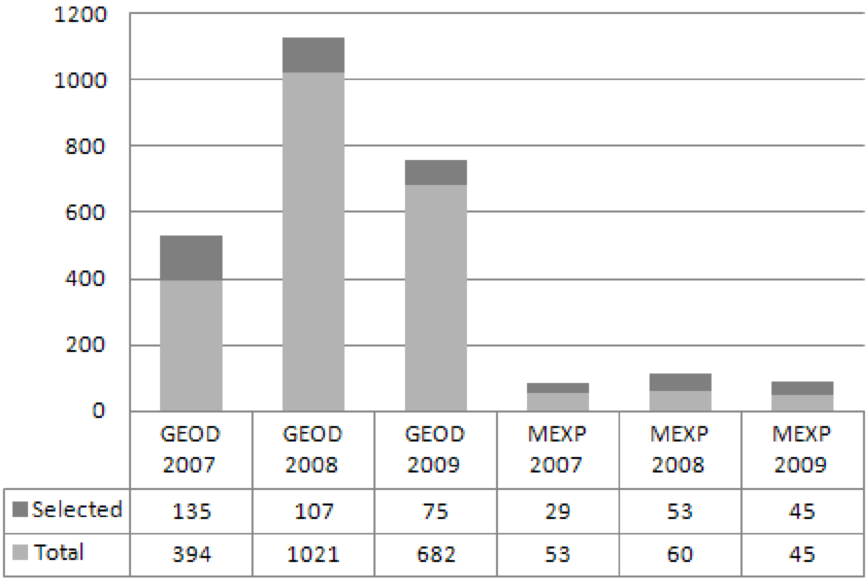


Repository	#Experiments
GEO (NCBI)	16639
ArrayExpress (EBI)	10571
SMD (Stanford)	496
caArray (NCBI)	156
CIBEX (Japan)	86

(a) Published experiments per repository.

(b) ArrayExpress evolution.

Fig. 1: Microarray experiments repositories overview.



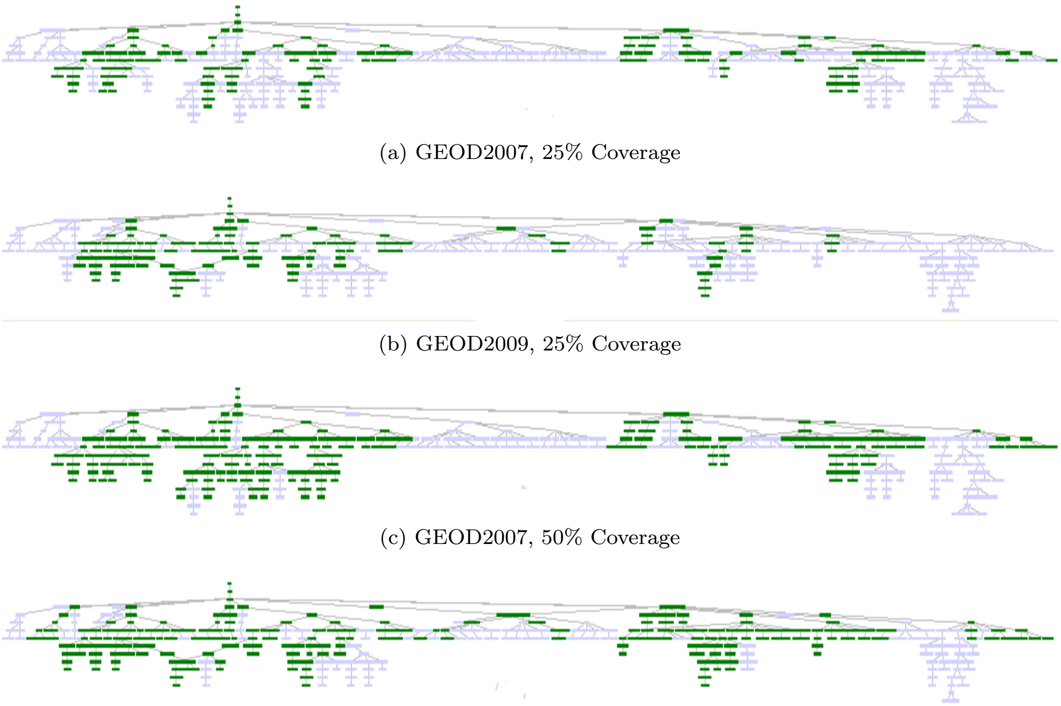
	GEOD 2007	GEOD 2008	GEOD 2009	MEXP 2007	MEXP 2008	MEXP 2009
■ Selected	135	107	75	29	53	45
■ Total	394	1021	682	53	60	45

Fig. 2: Number of selected experiments.

4. MAGE-ML SCHEMA USAGE AND EVOLUTION

Since most elements in the MAGE-ML schema are optional, usage patterns may vary substantially among data providers. In order to explore usage patterns we use DescribeX to process the structure of the selected experiments, allowing us to visually asses differences among data sources and usage evolution through years. For each data source presented in Section 3.1 we show the results of increasing coverage from 25% to 50%.

Figure 3 displays summary graphs for experiments imported from GEO into ArrayExpress, while Figure 4 shows summary graphs for experiments originally submitted to ArrayExpress. Several issues may be observed in these figures:



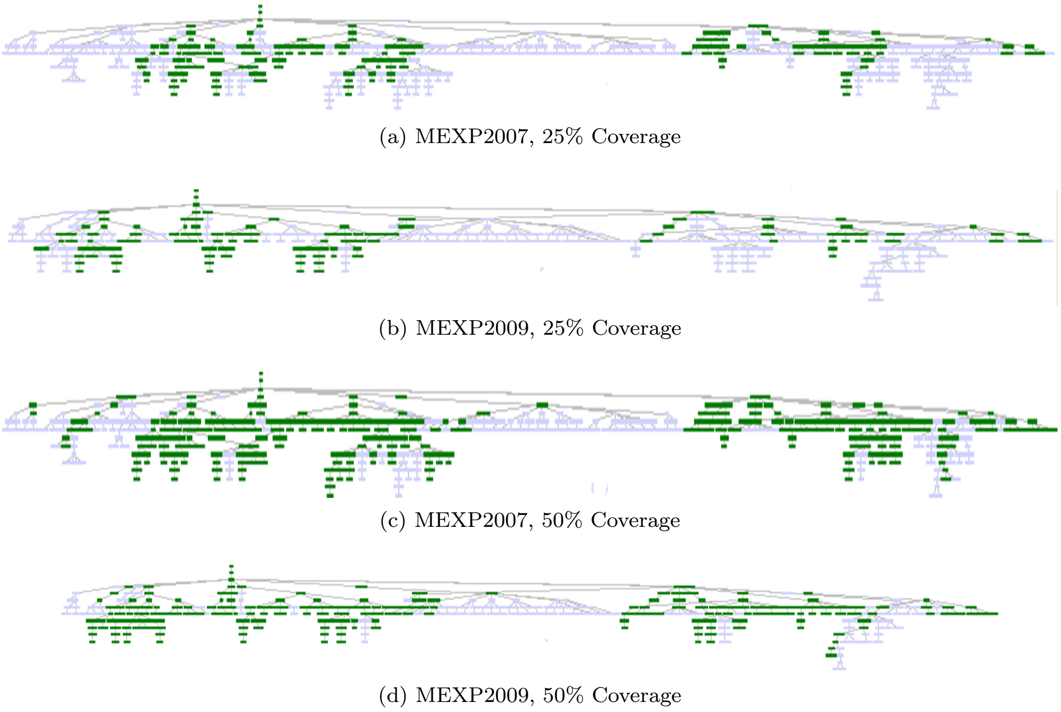
(a) GEOD2007, 25% Coverage

(b) GEOD2009, 25% Coverage

(c) GEOD2007, 50% Coverage

(d) GEOD2009, 50% Coverage

Fig. 3: Schema usage evolution (GEO imported experiments).



(a) MEXP2007, 25% Coverage

(b) MEXP2009, 25% Coverage

(c) MEXP2007, 50% Coverage

(d) MEXP2009, 50% Coverage

Fig. 4: Schema usage evolution (ArrayExpress original experiments).

—**Each data source uses different elements of the schema.** Consider for instance differences between Figure 3d and Figure 4d.

—For each data source, **schema usage seems to evolve through years**. Consider for example differences between Figure 3c and Figure 3d.

—**Schema usage seems to evolve independently for each data source.** For example, differences between Figure 3c and Figure 3d are not the same than between Figure 4c and Figure 4d.

From these issues, we conclude that MAGE-ML schema usage is not uniform, leading to differences in the quality of data, specially its completeness and accuracy. These usage patterns could be analyzed in depth by the microarray community and could be used, for example, to discover different in-lab practices that may lead to refinements and modifications of the MAGE-ML standard.

5. USE OF CONTROLLED VOCABULARY

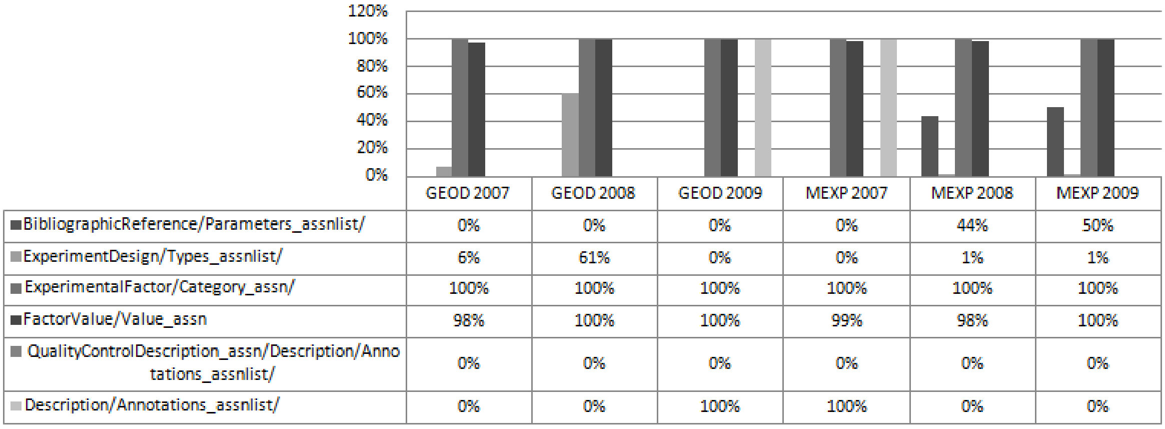
In this section we show accuracy measurements in the use of controlled vocabulary. While MAGE-ML provides a mechanism to standardize data representation for data exchange, a common terminology for data annotation is needed to support these standards. The MGED Ontology [Whetzel 2006] provides terms for annotating all aspects of a microarray experiment from experiment design of the experiment and array layout, through to preparation of biological samples and the protocols used to hybridize the RNA and analyze the data. An example is depicted in Figure 5, which represents an extract of MAGE-ML description file of experiment E-GEOD-4109 [4] obtained from ArrayExpress database. In this example a term in the ontology is used to specify one of the experimental protocols performed in the experiment.

```
<Protocol_package>
  <Protocol_assnlist>
    <Protocol identifier="P-G4109-1" text="no treatment"
                    name="treatment">
      <Type_assn>
        <OntologyEntry category="ProtocolType"
                  value="specified_biomaterial_action"/>
      </Type_assn>
    </Protocol>
  </Protocol_assnlist>
</Protocol_package>
```

Fig. 5: Example of use of *OntologyEntry* elements.

Within the MAGE-ML standard, *OntologyEntry* elements are aimed to contain references to individuals in MGED Ontology, but sometimes their references are incorrect. In [Etcheverry et al. 2010] we have introduced a simple method based on XML transformations to evaluate syntactic correcteness. This method replaces each *OntologyEntry* element with new elements according to the quality evaluation result. In this case each element has been replaced by an *OntologyEntryCorrect* or an *OntologyEntryIncorrect* element. Figure 6 presents measurement results in the *Experiment* subtree for each collection presented in Section 3.1, where some interesting issues may be noticed. For some ontology terms there are not clear trends in its accuracy evolution. For example, *ExperimentDesignTypes_ assnlist* references vary within GEO data source from 94% valid references in 2007 to 39% valid references in 2008 and 100% valid references in 2009. Other elements show a consistent pattern

[4] http://www.ebi.ac.uk/microarray-as/ae/browse.html?keywords=e-geod-4109



	GEOD 2007	GEOD 2008	GEOD 2009	MEXP 2007	MEXP 2008	MEXP 2009
■BibliographicReference/Parameters_assnlist/	0%	0%	0%	0%	44%	50%
■ExperimentDesign/Types_assnlist/	6%	61%	0%	0%	1%	1%
■ExperimentalFactor/Category_assn/	100%	100%	100%	100%	100%	100%
■FactorValue/Value_assn	98%	100%	100%	99%	98%	100%
■ QualityControlDescription_assn/Description/Annotations_assnlist/	0%	0%	0%	0%	0%	0%
■Description/Annotations_assnlist/	0%	0%	100%	100%	0%	0%

Fig. 6: Percentage of invalid ontology references in *Experiment* subtree.

of usage. For instance, experiments imported from GEO does not include references to *QualityControlDescription_ assn/Description/Annotations_ assnlist/* whereas all the references to this element within experiments originally submitted to ArrayExpress (MEXP experiments) are valid.

6. CONCLUSION

In this article, we extended previous results using DescribeX to asses data quality in microarray experiments. We showed how DescribeX helps to process and visualize large-scale collections of XML data, aiding in the process of pattern usage discovery. Even though is not always possible to quantitatively asses data quality in this context, our work shows that publicly available experimental data presents significant differences in its structure and content.

REFERENCES

Ali, M. S., Consens, M. P., Khatchadourian, S., and Rizzolo, F. DescribeX: Interacting with AxPRE Summaries. In *Proceedings of IEEE International Conference on Data Engineering*. Cancún, México, pp. 1540–1543, 2008.

Ali, M. S., Consens, M. P., and Rizzolo, F. Visualizing structural patterns in web collections. In *Proceedings of International World Wide Web Conference*. Banff, Canada, pp. 1333–1334, 2007.

Allison, D. B., Cui, X., Page, G. P., and Sabripour, M. Microarray Data Analysis: From Disarray to Consolidation and Consensus. *Nature Reviews Genetics* 7 (1): 55–65, Jan, 2006.

Bian, X. e. a. Data Submission and Curation for caArray, a Standard Based Microarray Data Repository System. In *Nature Precedings, 3rd International Biocuration Conference*, 2009.

Brazma, A. e. a. ArrayExpress: a Public Repository for Microarray Gene Expression Data at the EBI. *Nucl. Acids Res.* 31 (1): 68–71, January, 2003.

Brettschneider, J., Collin, F., Bolstad, B. M., and Speed, T. P. Quality Assessment for Short Oligonucleotide Microarray Data. *Technometrics* 50 (3): 241–264, August, 2008.

Consens, M. P., Rizzolo, F., and Vaisman, A. A. AxPRE Summaries: Exploring the (Semi-) Structure of XML Web Collections. In *Proceedings of IEEE International Conference on Data Engineering*. Cancún, México, pp. 1519–1521, 2008.

Demeter, J. e. a. The Stanford Microarray Database: Implementation of New Analysis Tools and Open Source Release of Software. *Nucl. Acids Res.* vol. 35, pp. D766, January, 2007.

Do, H. H., Kirsten, T., and Rahm, E. Comparative Evaluation of Microarray-based Gene Expression Databases. In *Proceedings of Fachtagung Datenbanksysteme für Business, Technologie und Web*. Leipzig, Germany, pp. 482–501, 2003.

Edgar, R., Domrachev, M., and Lash, A. E. Gene Expression Omnibus: NCBI Gene Expression and Hybridization Array Data Repository. *Nucl. Acids Res.* 30 (1): 207–210, 2002.

Etcheverry, L., Khatchadourian, S., and Consens, M. Quality Assessment of MAGE-ML Genomic Datasets Using DescribeX. In *Proceedings of International Conference on Data Integration in the Life Sciences*. Gothenburg, Sweden, pp. 192–206, 2010.

Goldman, R. and Widom, J. DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases. In *Proceedings of International Conference on Very Large Data Bases*. Athens, Greece, pp. 436–445, 1997.

Gray, J., Liu, D., Santisteban, M., Szalay, A., DeWitt, D., and Heber, G. Scientific Data Management in the Coming Decade. *SIGMOD* vol. 34(4), pp. 34–41, 2005.

Ikeo, K., Ishi-i, J., Tamura, T., Gojobori, T., and Tateno, Y. CIBEX: Center for Information Biology gene EXpression database. *Comptes Rendus Biologies* 326 (10-11): 1079 – 1082, 2003.

Kohane, I. S., Kho, A., and Butte, A. J. *Microarrays for an Integrative Genomics*. MIT Press, 2002.

Lockhart, D. J. and Winzeler, E. A. Genomics, gene expression and DNA arrays. *Nature* 405 (6788): 827–836, June, 2000.

Madnick, S. E., Wang, R. Y., Lee, Y. W., and Zhu, H. Overview and Framework for Data and Information Quality Research. *Journal of Data and Information Quality* 1 (1): 1–22, 2009.

Martinez, A. and Hammer, J. Making Quality Count in Biological Data Sources. In *Proceedings of International Workshop on Information Quality in Information Systems*. Baltimore, USA, pp. 16–27, 2005.

Missier, P., Embury, S. M., Greenwood, M., Preece, A. D., and Jin, B. Managing Information Quality in E-science: the Qurator Workbench. In *Proceedings of ACM SIGMOD International Conference on Management of Data*. Beijing, China, pp. 1150–1152, 2007.

Müller, H. and Naumann, F. Data Quality in Genome Databases. In *Proceedings of International Conference on Information Quality*. Cambridge, MA, USA, pp. 269–284, 2003.

Parkinson, H. e. a. ArrayExpress update–from an archive of functional genomics experiments to the atlas of gene expression. *Nucl. Acids Res.* 37 (Database issue): D868–872, January, 2009.

Rayner, T. F. e. a. MAGETabulator, a suite of tools to support the microarray data format MAGE-TAB. *Bioinformatics* 25 (2): 279–280, January, 2009.

Samavi, R., Consens, M., Khatchadourian, S., and Topaloglou, T. Exploring PSI-MI XML Collections Using DescribeX. *Journal of Integrative Bioinformatics* 4 (3): 70, 2007.

Spellman, P. T. e. a. Design and Implementation of Microarray Gene Expression Markup Language (MAGE-ML). *Genome Biology* 3 (9): research0046.1–research0046.9, August, 2002.

Whetzel, P. L. e. a. The MGED Ontology: a Resource for Semantics-based Description of Microarray Experiments. *Bioinformatics* 22 (7): 866–873, April, 2006.