FCAWarehouse, a prototype online data repository for FCA

Constantinos Orphanides and George Georgiou

Conceptual Structures Research Group
Communication and Computing Research Centre
Faculty of Arts, Computing, Engineering and Sciences
Sheffield Hallam University, Sheffield, UK
`c.orphanides@shu.ac.uk` `georgios.georgiou@student.shu.ac.uk`

Abstract. This paper presents FCAWarehouse, a prototype online data repository for FCA. The paper explains the motivation behind the development of FCAWarehouse and the features available, such as the ability to donate datasets and their respective formal contexts, the ability to generate artificial formal contexts on-the-fly, and how these features are also available through a set of web-services. The paper concludes by suggesting future work in order to enhance it's usability.

1 Introduction

In recent years, different means of archiving empirical data have been considered and implemented to build a space where data collection could be exploited and analyzed accordingly, if required. The different archiving techniques not only introduced communities an alternative approach of maintaining the quality and accessibility of data, but also improved the indexing of available data collections with proper categorization. An example of such an archiving technique are online data repositories.

Online data repositories are digital libraries which allow the comprehensive collection, management and preservation of digital content and the ability to offer it to targeted user communities [10,7,8]. An example of such a repository is the UCI Machine Learning Repository, a repository used by the machine learning community for the empirical analysis of machine learning algorithms [6]. Data repositories exist for numerous targeted communities; however, a centralized, public resource of data in FCA formats for the FCA community has not been implemented to date.

The idea of an FCA online data repository, to provide FCA practitioners with a resource of public datasets in FCA formats, was proposed in [2] at 2009. The proposal envisioned the automatic conversion of uploaded traditional datasets into formal contexts, as well as the ability to generate artificial datasets in CSV format, to be then converted into a formal context with the user being able to determine the number of objects, number of attributes and the density. The automatic conversion of datasets to formal contexts would be handled by a

"Data-to-FCA" converter and the tool FcaStone[1] would be used to convert formal contexts from one FCA format to another. Both of these tools would be incorporated in the repository as components. The architecture of the proposed repository is shown in Figure 1.

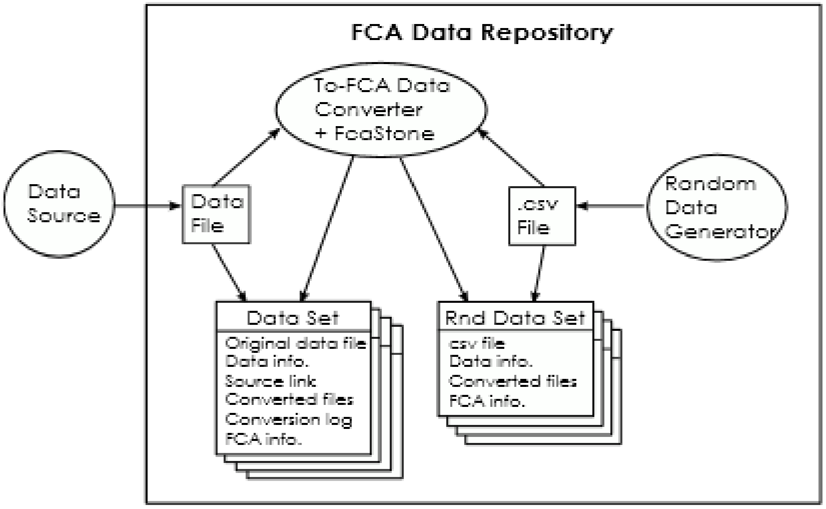


Fig. 1. Proposed FCA Data repository system architecture in [2].

The "Data-to-FCA" converter of the proposed repository was the most complex component to implement with regards to the effort and time needed to develop it. However, since the publication of the aforementioned paper, the "Data-to-FCA" converter has been developed as a standalone desktop application called FcaBedrock[2], a formal context creator for FCA with the ability of converting datasets in various formats to formal contexts in the Burmeister (`.cxt`) or FIMI (`.dat`) formats [5,4]. The formal contexts generated by FcaBedrock can be then loaded in In-Close[3], a fast formal concept miner, to count the number of formal concepts in a formal context and produce, if necessary, smaller sub-contexts based on the well-known notion of *minimum support* [3,1]. Subsequently, the motivation behind implementing an FCA data repository became stronger and resulted in the development of a prototype data repository for FCA, named FCAWarehouse.

[1] `http://sourceforge.net/projects/fcastone`
[2] `http://sourceforge.net/projects/fcabedrock`
[3] `http://sourceforge.net/projects/inclose`

2 FCAWarehouse

FCAWarehouse[4] is a prototype data repository for FCA developed as part of a BSc Computing final year project [9] at Sheffield Hallam University (SHU). It provides the ability of donating datasets and their respective formal contexts, browsing and searching datasets, an administration back-end for librarians to manage donated datasets, creating artificial formal contexts, as well as providing all of it's functionality through a set of ReSTful[5] web services to make FCAWarehouse interoperable with third-party FCA tools. A diagram of it's architecture is shown in Figure 2 and an explanation of it's features is given in the following sections.

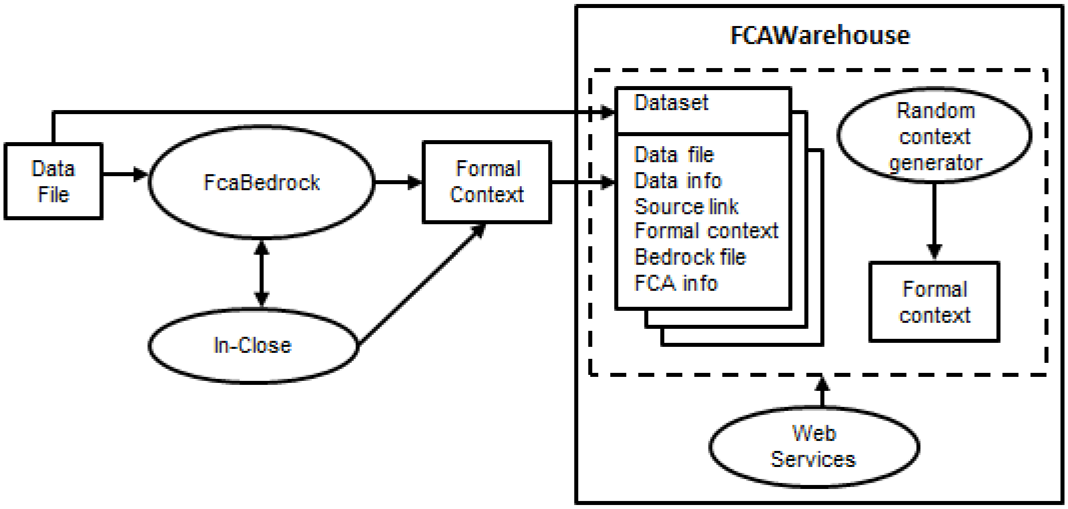


Fig. 2. System architecture of FCAWarehouse.

2.1 Datasets

Each dataset in FCAWarehouse is comprised of metadata such as name, additional information, original source as well as the original data and it's respective formal context (Figure 3).

All of the files and metadata are provided by the donator during the donation process. In cases where the formal contexts are created using FcaBedrock, a

[4] http://www.fcawarehouse.com

[5] Representational State Transfer (ReST). http://www.ibm.com/developerworks/webservices/library/ws-restful/

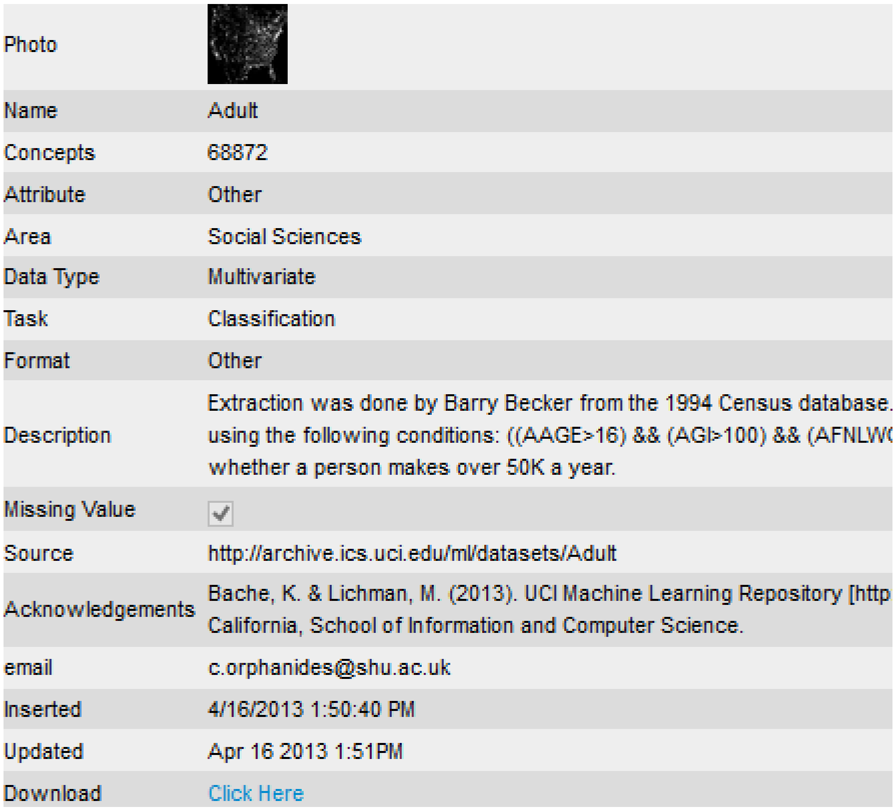


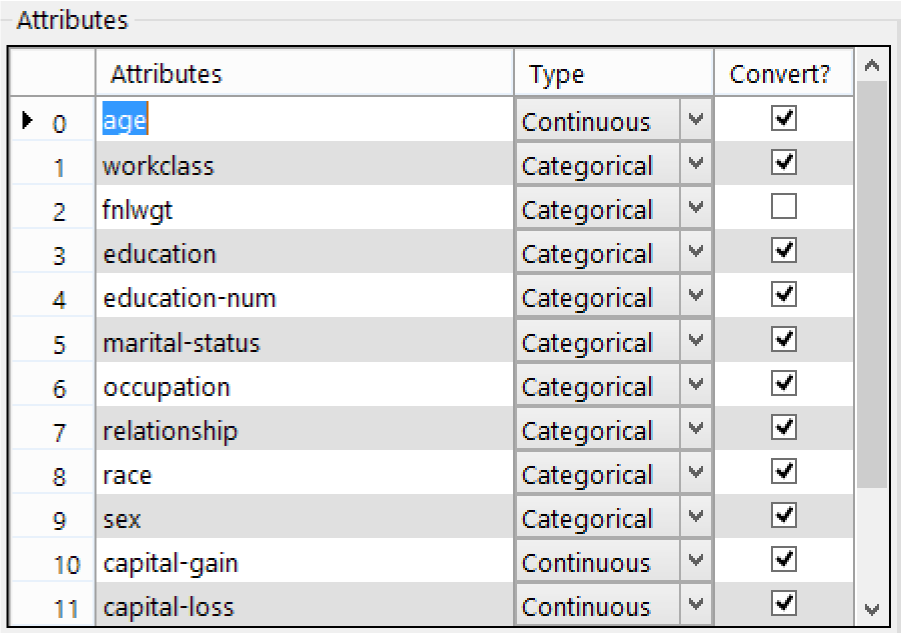
Photo	
Name	Adult
Concepts	68872
Attribute	Other
Area	Social Sciences
Data Type	Multivariate
Task	Classification
Format	Other
Description	Extraction was done by Barry Becker from the 1994 Census database. using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLW whether a person makes over 50K a year.
Missing Value	☑
Source	http://archive.ics.uci.edu/ml/datasets/Adult
Acknowledgements	Bache, K. & Lichman, M. (2013). UCI Machine Learning Repository [http California, School of Information and Computer Science.
email	c.orphanides@shu.ac.uk
Inserted	4/16/2013 1:50:40 PM
Updated	Apr 16 2013 1:51PM
Download	Click Here

Fig. 3. Example of a dataset entry in FCAWarehouse (partial screenshot).

Bedrock file (`.bed`) is also provided which contains the metadata of the conversion so that users can load the original data and the Bedrock file in FcaBedrock and have a look at, or modify, the parameters and conversion criteria set for each attribute in the original dataset (Figure 4).

The files donated for each dataset are stored physically on the server, with the database only holding the URL to each dataset. As the prototype is currently hosted on a server with limited storage capabilities, all files uploaded for a dataset during the donation process are automatically compressed as `.zip` files.

After a user has donated a dataset, the librarians are notified about the new submission and have the ability of accepting or declining the submitted dataset. Consequently, only datasets marked as 'Accepted' by the librarians are visible to end-users.



	Attributes	Type	Convert?
0	age	Continuous	☑
1	workclass	Categorical	☑
2	fnlwgt	Categorical	□
3	education	Categorical	☑
4	education-num	Categorical	☑
5	marital-status	Categorical	☑
6	occupation	Categorical	☑
7	relationship	Categorical	☑
8	race	Categorical	☑
9	sex	Categorical	☑
10	capital-gain	Continuous	☑
11	capital-loss	Continuous	☑

Fig. 4. FCAWarehouse's Bedrock (`.bed`) file of the Adult dataset loaded in FcaBedrock (partial screenshot).

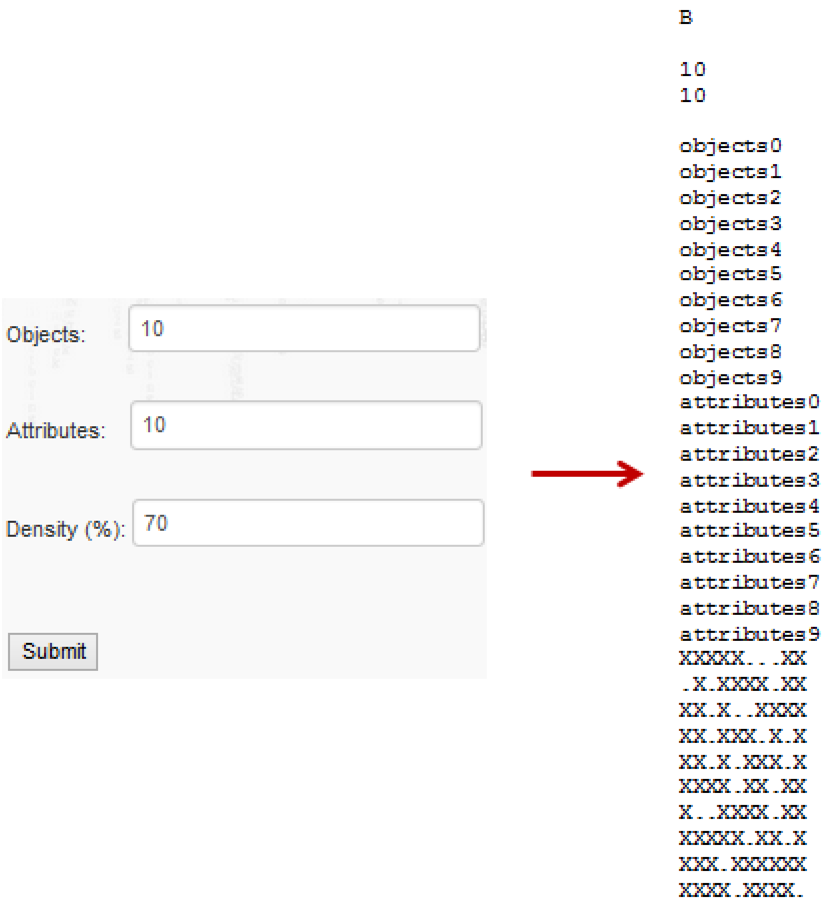
2.2 Artificial formal contexts

FCAWarehouse provides the ability of generating artificial formal contexts in the Burmeister (`.cxt`) format by predefining the number of formal objects, number of formal attributes and the density. Developers of tools such as the formal concept miners In-Close[6] and FCbO[7] can use this feature to generate formal contexts of size of their choice in order to test the limits and efficiency of their algorithms.

The generated artificial formal contexts are randomly generated by predefining the number of objects, number of attributes and density of the formal context. The density is defined as a percentage indicating the minimum amount of crosses (formal attributes) that each row (formal object) should contain in the formal context. Setting for example number of objects to 10, number of attributes to 10 and density percentage to 70% will result in a formal context with 10 formal objects and 10 formal attributes where *at least* 70% of each row is consisted of crosses (Figure 5). Setting the density to 0 will generate a random artificial formal context with no density criteria.

[6] `http://sourceforge.net/projects/inclose`

[7] `http://sourceforge.net/projects/fcalgs`



Objects: 10

Attributes: 10

Density (%): 70

```
B

10
10

objects0
objects1
objects2
objects3
objects4
objects5
objects6
objects7
objects8
objects9
attributes0
attributes1
attributes2
attributes3
attributes4
attributes5
attributes6
attributes7
attributes8
attributes9
XXXXX...XX
.X.XXXX.XX
XX.X..XXXX
XX.XXX.X.X
XX.X.XXX.X
XXXX.XX.XX
X..XXXX.XX
XXXXX.XX.X
XXX.XXXXXX
XXXX.XXXX.
```

Fig. 5. Example of an artificial formal context with 10 objects, 10 formal attributes and density set to 70%.

2.3 Web services

FCAWarehouse implements a set of web services for the interoperability of FCAWarehouse with third-party FCA applications. At the moment there are a total of four web services, namely:

- **GenerateContext**: Accepts as input the number of objects, number of attributes, density percentage and outputs a corresponding artificial Burmeister formal context in XML.
- **GetAllDataSets**: Returns all datasets, along with their metadata, in XML. Only datasets marked as ‘Accepted’ by the librarians are returned.

- **Get10MostRecentDatasets**: Returns the 10 most recent datasets along with their metadata, in XML. Only datasets marked as 'Accepted' by the librarians are returned.
- **SearchDataset**: Accepts a string as input and returns any datasets which contain the given string in their name, along with their metadata, in XML. Only datasets marked as 'Accepted' by the librarians are returned.

3 Further Work and Conclusion

While useful in it's current state, FCAWarehouse is still in a prototypical state; a number of improvements can be implemented to enhance it's usability. For example, the feature of artificially generating formal contexts can be extended to simulate formal contexts that real datasets would produce. This can be achieved by defining sets of adjacent columns in the formal context to represent a single attribute. Assuming, for example, a categorical attribute (with each of it's values being a formal attribute in the formal context) and each formal object only being able to have one of it's values (quite common in real datasets), will result in a formal context where no more than one cross will exist in the columns representing that attribute. The same logic can be applied to various type of attribute: boolean attributes could be interpreted as single formal attributes in the formal context and continuous attributes could be grouped using ranges rather than one formal attribute for each numerical value. In this way, FCAWarehouse can act as a benchmarker for the comparison of tools and algorithms by providing citable random data as well as converted real datasets.

A feature which could prove quite useful in the future, with some modifications and adjustments, are FCAWarehouse's web services. Interesting use-cases can emerge from this feature; for example, the formal context creator FcaBedrock could feature a "Create formal context from FCAWarehouse" option. By using the provided web services, a list of the available datasets could appear in FcaBedrock. Selecting one of the datasets could automatically download the dataset and start auto detecting its values to create a formal context. Creating a formal context in FcaBedrock requires inputting a dataset, defining its metadata and then creating the formal context; considering, however, the "Generate Context" web service, the ability of creating artificial formal contexts in FcaBedrock without requiring initial data as input could be made possible with minimal effort.

Further work also includes the incorporation of the tool FcaStone, to convert data between FCA and non-FCA formats, in FCAWarehouse to offer the power of FCA to those currently outside of the FCA community.

The final vision of FCAWarehouse is of an online FCA data repository which facilitates the creation, conversion and donation of datasets for FCA, providing a useful collection of real and artificial datasets in a wide variety of FCA and non-FCA formats to open the way for the wider use of FCA.

References

1. Andrews, S. (2011). *In-Close2, a High Performance Formal Concept Miner*. In: Proceedings of the 19th International Conference on Conceptual Structures (ICCS) 2011. LNAI 6828. Berlin: Springer-Verlag. pp. 50–62.
2. Andrews, S. (2009a). *Data Conversion and Interoperability for FCA*. In: Proceedings of the 4th Conceptual Structures Tool Interoperability Workshop (CS-TIW) 2009, held in conjunction with the 17th International Conference on Conceptual Structures (ICCS) 2009: "Leveraging Semantic Technologies". pp. 42–49.
3. Andrews, S. (2009b). *In-Close, a Fast Algorithm for Computing Formal Concepts*. In: Rudolph, Dau, Kuznetsov (Eds.): Supplementary Proceedings of ICCS'09, CEUR WS 483.
4. Andrews, S. and Orphanides, C. (2012). *Knowledge Discovery Through Creating Formal Contexts*. In: International Journal of Space-Based and Situated Computing, **2**(2). pp. 123–138.
5. Andrews, S. and Orphanides, C. (2010). *FcaBedrock, a Formal Context Creator*. In: "Conceptual Structures: From Information to Intelligence", Proceedings of the 18th International Conference on Conceptual Structures (ICCS) 2010, LNAI 6208. pp. 181–184.
6. Bache, K. and Lichman, M. (2013) *UCI Machine Learning Repository* [`http://archive.ics.uci.edu/ml`]. Irvine, CA: University of California, School of Information and Computer Science.
7. Candela, L., Athanasopoulos, G., Castelli, D., El Raheb, K., Innocenti, P., Ioannidis, Y., Katifori, A., Nika, A., Vullo, G. and Ross, S. (2011) *The Digital Library Reference Model*. Last accessed 01 April 2013 at: `http://bscw.research-infrastructures.eu/pub/bscw.cgi/d222816/D3.2b%20Digital%20Library%20Reference%20Model.pdf`
8. DSpace Cambridge (2012). *Digital Repositories*. [online]. Last accessed 18 April at `http://www.lib.cam.ac.uk/dataman/pages/repositories.html`
9. Georgiou, G. (2013). *FCAWarehouse, an Online FCA Data Repository*. BSc (Hons) Computing final year project, Sheffield Hallam University, Sheffield, UK, 2013.
10. Scherle, R., Carrier, S., Greenberg, J., Lapp, H., Thompson, A., Vision, T. and White, H. (2008). *Building Support for a Discipline-Based Data Repository*. In: Third International Conference on Open Repositories 2008, Southampton. Last accessed 20 March 2013 at: `http://pubs.or08.ecs.soton.ac.uk/35/1/submission_177.pdf`