MONITORING FILE FORMAT OBSOLESCENCE IN REPOSITORIES

An applied method

Sam Alloing
National Library of the Netherlands
Netherlands
sam.alloing@kb.nl
0000-0002-1254-1483

Abstract – The Dutch Digital Heritage Network (DDHN) wants to improve the monitoring of file format obsolescence. The Preservation Watch group researched on how institutions can monitor the life cycle of file formats in their repositories and how the monitoring could be implemented on a broader scale. Monitoring file format life cycle implies there needs to be a way to measure format obsolescence or helps an institution to identify when a file format is getting obsolete. The applied research identified the needed information and used a known model to search for trends and is applied in widespread areas. The model was compared with a naive method to evaluate the more complex method. This approach was tested in different types of repositories and used different file formats to research the robustness of the approach. This paper will investigate the possibilities and shortcomings of this method and further research that is required.

Keywords – preservation watch, file formats, applied research, file format obsolescence, Bass diffusion model

Conference Topics – From Theory to Practice

I. INTRODUCTION

Format obsolescence is a widely discussed topic in digital preservation. There are different strategies dealing with obsolete file formats like file format migration or emulation. The moment when to execute the preservation strategy is not an easy decision. Some policies use a late migration strategy. This strategy needs information about when to take a preservation action, so the migration is not too late and files can still be opened.

This paper uses the outcome of an earlier paper [1] that investigated the Bass Diffusion Model as a possible solution for detecting file format obsolescence and builds upon the results by using repositories of different institutions. The Bass Diffusion Model is used in a wide variety of use cases and is not specific for digital preservation.

II. METHODOLOGY

The increase and decline of products is described and predicted in the Bass Diffusion Model [2]. The model describes the life cycle of a product where innovators are early adopters of a product and later the imitators join with the big increase in use and the diminishing effect of laggers that follow after that. This gives the curve a typical bell shape with a steep start and a long tail [3]. Depending on where a product is in its life cycle it shows a cut out of the bell shape.

This model was also previously applied in the area of file format obsolescence and deemed useful. The model was applied in a context of a web archive and this has limitations on which file formats can be researched. The repositories of an institution are also different then the corpus of a web archive, because the last one has predominantly file formats that are used on the web, like for example HTML[4].

To help the interpretation of the output of the Bass Diffusion model a second model was used as a

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

https://hdl.handle.net/2142/121117

reference model. The linear regression model is used, because it is a simple model that represents the naive approach. To be useful the more complex model needs to be a better explanation then the simpler model otherwise there is no added value. Because the simplicity of the linear regression, it is only applied on file formats with declining popularity. This way linear regression can be used as an evaluation model.

The aim of the research is also to look into the prediction capabilities of the models and we use the last quarters as a test set for prediction. In the plots it is shown as a green (Bass test) and purple line (Lineair test). Most of the data is used as a training set and the smaller test set the model needs to predict the course of the life cycle of the file format. This is used as an indication of the reliability of the prediction by the different models. The Blue line on the plot is the number of files.

III. USED APPROACH

The approach looks for diminishing delivery to an institution or use on the internet of a file format over time. This is an indication used in the model as a diminishing popularity of a file format. The time period is over several years. Because there are also rare file formats the time period is over several quarters in a year so there are enough data points to make a predicting model.

The life cycle can include an increase and a decrease of popularity and shows at which stage in the life cycle a file format is. This also brings up the question if there is a threshold which indicates if a format is getting obsolete or if the file format monitoring can be automated. Not only the monitoring of a single file format is investigated, but also if file formats are linked together and if a file format is a predecessor which shows a decrease in popularity and if there is a successor that shows an increase in popularity .

A last and final factor that is important to monitor is the relation between applications and file formats. A decrease in the number of applications that can open or write a certain file format over time gives an indication of a file format becoming obsolete, because a decrease in popularity of a file format doesn't need to mean obsolescence. The combination of file formats and applications will be used as an indication of obsolescence.

IV. DATA QUALITY

For the analysis we used two types of data, data from Common Crawl and data from different institutional repositories (Netherlands Institute for Sound and Vision and Data Archiving and Networked Services (DANS)). The Common Crawl data is publicly available data from internet crawls. There are summaries available of for example mime type and this prevents the need to process all the Common Crawl dataset [5]. The mime types are identified by Apache Tika [6]. The summarized metadata was used for this analysis. The disadvantage is that only date from 2017 onwards was usable, but in general this is also the year in which Common Crawl data is more reliable [7]. This limits the results of the output as an indication of the file format lifecycle on the broader and international scale.

This is a recurring theme, getting usable information is a challenge also for repositories of institutions. Although the information seems simple, just the date that indicates the creation of the file or a substitute date of the resource that is preserved like publication date. But that was a challenge. Most institutions were able to produce ingest dates, but that isn't a sufficient date, for example because of migrations of content when changing systems. The other challenge was to produce file format identification that was precise enough. Most institutions could only produce mime types or file extensions which don't have file format version information. For example the MS Access database format, MDB, can contain a wide range of MS Access database software versions [8]. Institutions using for example the Pronom PUID which can describe the MS Access file format much more precise [9].

V. ANALYSIS AND RESULTS

A. Common Crawl

The Common Crawl analysis was used to test the approach on a large scale data set. The hypothesis is to use this as a comparison to the repository level and use this as an extra evaluation criteria to interpret the results of the repository level.

Of the different analyzed mime types XHTML and GIF will be discussed [10].

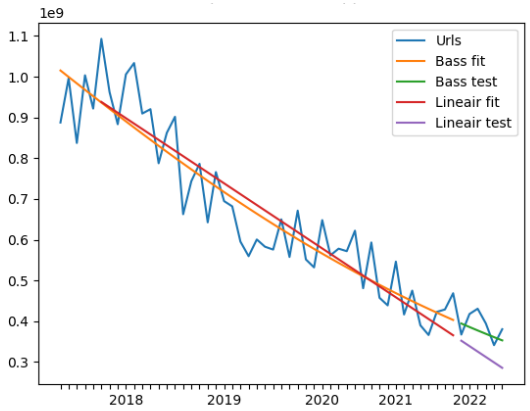

Fig 1 XHTML plot. 2022 CC-BY-SA-4.0 Rein van 't Veer/DDHN

The XHTML plot (Fig 1) shows a declining graph. The format is in the downwards spiral of the bell shaped curve, but is not in danger of getting obsolete. The format still constitutes 12% of the billions of pages harvested by Common Crawl, so no obsolescence is expected. Browsers can still open the file format as well.

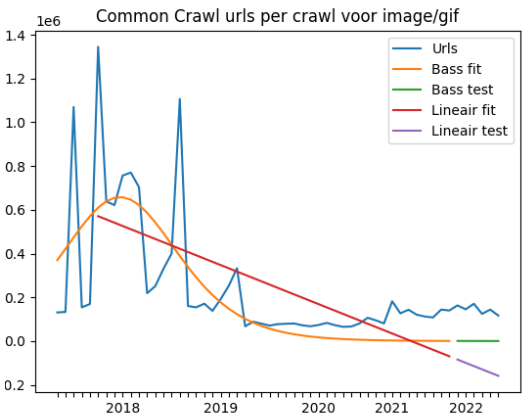

Fig 2 GIF plot. 2022 CC-BY-SA-4.0 Rein van 't Veer/DDHN

The GIF file format (Fig 2) is also in a decline, but there is a part that shows an increase. The increase is due to the incomplete set. The file format is already in use for a long time, but the limited time period of the data set and the erratic peaks throw off the Bass model. The linear regression just shows a decline, but the prediction goes below 0. The Bass model shows a more realistic trajectory.

Of the 26 investigated formats in the Common Crawl data set, the Bass model had in 13 cases a better prediction than the linear regression. In 3 cases the accurateness was the same and in 8 cases the linear regression performed better. The reason for these errors is probably comparable to the GIF case already discussed, the erratic peaks. This is also suggested by the other plots of other data sets.

B. Data Archiving and Networked Services (DANS)

DANS [11] is an institute in the Netherlands that preserves scientific data from scientific institutes. The data set from the archaeological repository is used. In this data set the analysis of multiple linked file formats was possible. The Microsoft Office formats (MS Word and Excel) show a linked file format lifecycle between different formats. The case of MS Excel formats XLS (Fig 3) and XLSX (Fig 4) is discussed as it shows the evolution clearly.

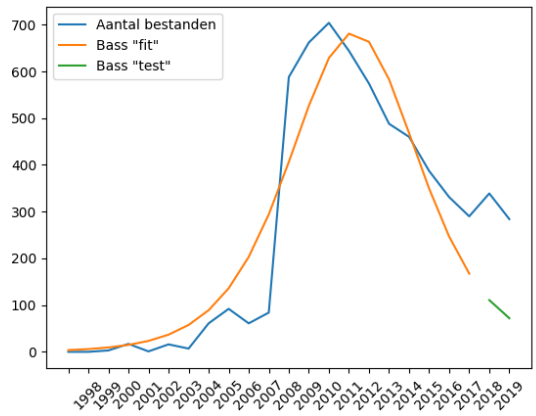

Fig 3 XLS plot. 2022 CC-BY-SA-4.0 Rein van 't Veer/DDHN

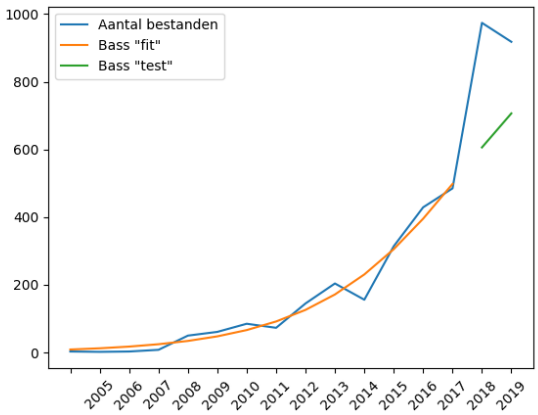

Fig 4 XLSX plot. 2022 CC-BY-SA-4.0 Rein van 't Veer/DDHN

These two plots show the decline of the XLS format and around the same time an increase of the XLSX format. This is expected as the XLS format is an older format that has been gradually phased out by Microsoft in favor of XLSX [12]. Microsoft Access doesn't show this trend in the DANS repository, the MDB file format (Fig 5) which is older than the ACCDB file format (Fig 6) still is very popular and is still increasing. This is unexpected, but can be explained by the specific case of archaeological data sets where MS Access is popular software and database templates in MDB file format are used and reused over and over. Also the number of files received is

much lower for the ACCDB file format (Blue line). This throws off the Bass Model prediction with a sharp increase in ACCDB and a decrease in the case of MDB.

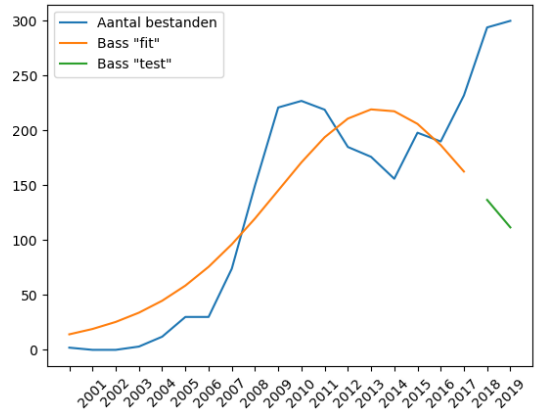

Fig 5 MDB plot. 2022 CC-BY-SA-4.0 Rein van 't Veer/DDHN

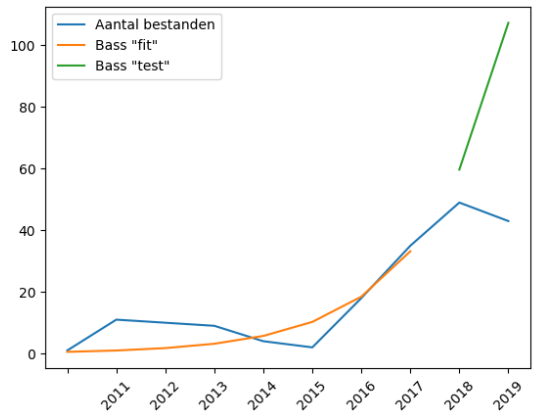

Fig 6 ACCDB plot. 2022 CC-BY-SA-4.0 Rein van 't Veer/DDHN

The DANS data set contains more file formats, these are not discussed here, but are described in an article [13].

VI. Application as extra dimension

It becomes apparent by the DANS data that file format is not the only dimension to look at when analyzing the file format life cycle. There is a need for an extra dimension, what application still supports the file format.

To start this analysis a good source of information is necessary. Different possible sources have been researched: Guide of preferred file formats [14], NARA digital preservation framework [15], Wikidata [16] and Pronom [17]. To evaluate the sources the case of Microsoft Access was used. During the analysis of the results of the DANS data set the MDB file format showed declining support in Microsoft Access and is in danger of getting obsolete [18]. This case shows that there is fine grained information needed between file format and application. The data model of Pronom and Wikidata can store the information that is needed to support the research. The problem with Pronom is the information not kept up-to-date [19]. The Wikidata data model has the potential to support the connection between file format and application, but the link is not yet sufficiently provided. The application version information is a literal and not an entity. A literal is a string of information and is not easy to query or it is not possible to link information to a literal. This is all possible with an entity, but in the case of Microsoft Access, this is most of the time not available. For example MS Access file format version 95, has as software version identifier 95 [20]. Microsoft Access Database, version 2007 [21] is an entity and queries are possible of for example the number of applications that can read the file format [22]. This shows potential but needs to be researched more and more data needs to be added like for example discontinued date [23].

VII. Conclusion

The research shows that the Bass Model can be used as a method to evaluate the format obsolescence, but it is not an automated process because the results need to be interpreted and understood in the specific context of a repository or in the broader scale, due to the specific community the repository serves or due to data quality issues. The method helps with summarizing the file format information and gives insight in the life cycle of the file format. The relation between the broad internet scale data set and the repository level data sets needs more research because of limited data sets and different file formats researched.

The relation between file format and application needs to be researched more, certainly if the analysis needs to be combined with the file format information and help to improve the file format life cycle analysis.

VIII. Acknowledgement

This research was conducted by Rein van 't Veer of Antfield Creations and The Preservation Watch working group of The Dutch Digital Heritage Network and.

The Dutch Digital Heritage Network is formed by organizations in the fields of culture, heritage, education, and research together. With suppliers of

heritage software, provinces and municipalities we are working on the implementation of the National Strategy Digital Heritage, supported by the Ministry of Education, Culture and Science.

REFERENCES

[1] Duretec, K. and Becker, C. (2017), Format technology lifecycle analysis. Journal of the Association for Information Science and Technology, 68: 2484-2500. https://doi.org/10.1002/asi.23881

[2] Bass, Frank M. (1969), A New Product Growth for Model Consumer Durables, Management Science, Vol. 15, No. 5, Theory Series, p. 215-227 https://math.la.asu.edu/~dieter/courses/APM_598/Bass_69.pdf

[3] Bass diffusion model. https://en.wikipedia.org/wiki/Bass_diffusion_model

[4] See HTML in MIME Types. https://commoncrawl.github.io/cc-crawl-statistics/plots/mimetypes

[5] MIME Types. https://commoncrawl.github.io/cc-crawl-statistics/plots/mimetypes

[6] Apache Tika - a content analysis toolkit. https://tika.apache.org/

[7] Size of Common Crawl Monthly Archives. https://commoncrawl.github.io/cc-crawl-statistics/plots/crawlsize

[8] Microsoft Access MDB File Format Family. https://www.loc.gov/preservation/digital/formats/fdd/fdd000462.shtml

[9] Wikidata Query about file format with file extension MDB and PUID https://w.wiki/6QXs

[10] For more mime types see (Dutch only): Rein van 't Veer (2022) Monitoring van bestandsformaten 2: het internet als archief, het Bass-model in de praktijk: welke internetformaten zijn aan het verdwijnen? https://kia.pleio.nl/groups/view/4fc4e83a-f55b-4000-b1cb-3fe9a16d3f93/kennisplatform-preservation/blog/view/b4b724a2-0683-438d-897c-717b95a57071/monitoring-van-bestandsformaten-het-internet-als-archief-het-bass-model-in-de-praktijk-welke-internetformaten-zijn-aan-het-verdwijnen

[11] DANS | Centre of expertise & repository for research data. https://dans.knaw.nl/en/

[12] Library of Congress, XLSX Transitional (Office Open XML), ISO 29500:2008-2016, ECMA-376, Editions 1-5 https://www.loc.gov/preservation/digital/formats/fdd/fdd000398.shtml, see Adoption

[13] Rein van 't Veer (2022) Monitoring van bestandsformaten 4: formaten in gebruik bij Data Archiving and Networked Services (DANS) https://kia.pleio.nl/groups/view/4fc4e83a-f55b-4000-b1cb-3fe9a16d3f93/kennisplatform-preservation/blog/view/b92f7a1e-fd6a-4c88-875c-5bccc470554c/monitoring-van-bestandsformaten-formaten-in-gebruik-bij-data-archiving-and-networked-services-dans (Dutch only)

[14] DDHN, Summary Guide to Prefered Formats. https://www.wegwijzervoorkeursformaten.nl/index.php/Summary_Guide_to_Prefered_Formats

[15] NARA, U.S. National Archives and Records Administration Digital Preservation Framework. https://github.com/usnationalarchives/digital-preservation#the-nara-risk-and-prioritization-matrix

[16] Wikidata. https://www.wikidata.org/wiki/Wikidata:Main_Page

[17] PRONOM https://www.nationalarchives.gov.uk/PRONOM/

[18] Microsoft, Which Access file format should I use? https://support.microsoft.com/en-us/office/which-access-file-format-should-i-use-012d9ab3-d14c-479e-b617-be66f9070b41

[19] This is confirmed by Francesca Mackenzie, Digital Archivist at UK National Archives and responsible for Pronom.

[20] Wikidata, Microsoft Access Database, version 95. https://www.wikidata.org/wiki/Q48004869

[21] Wikidata, Microsoft Office Access 2007. https://www.wikidata.org/wiki/Q46049725

[22] Wikidata query, Applications that can read MS Access file formats https://w.wiki/6RMa

[23] Wikidata, discontinued date. https://www.wikidata.org/wiki/Property:P2669