Accepted for publication in the
Journal of the Association for Information Science and Technology (JASIST)

March 2017

Format Technology Lifecycle Analysis

Kresimir Duretec

Vienna University of Technology, Institute of Software Technology and Interactive Systems, Information and Software Engineering Group
Favoritenstrasse 9-11/188
1040 Vienna
Austria
email: kresimir.duretec@tuwien.ac.at

Christoph Becker

University of Toronto, Faculty of Information
140 St George Street
Toronto, ON M5S 3G6
Canada
Email: christoph.becker@utoronto.ca

Format Technology Lifecycle Analysis

Abstract

The lifecycles of format technology have been a defining concern for digital stewardship research and practice. However, little evidence exists to provide robust methods for assessing the state of any given format technology and describing its evolution over time.

This article introduces relevant models from diffusion theory and market research and presents a replicable analysis method to compute models of technology evolution. Data cleansing and the combination of multiple data sources enable the application of non-linear regression to estimate the parameters of the Bass diffusion model on format technology market lifecycles.

Through its application to a longitudinal dataset from the UK web archive, we demonstrate that the method produces reliable results and show that the Bass model can be used to describe format lifecycles. By analyzing adoption patterns across market segments, new insights are inferred about how the diffusion of formats and products such as applications occurs over time. The analysis provides a stepping stone to a more robust and evidence-based approach to model technology evolution.

Keywords

digital preservation, digital curation, digital stewardship, file formats, obsolescence, web archive, technology evolution

Introduction

The need to understand how technology evolves is a central concern of digital stewardship. The Open Archival Information Systems Reference Model (OAIS) defines the long-term aspect of digital preservation as "long enough to be concerned with the impacts of changing technologies, including support for new media and data formats, or with a changing user community" (Consultative Committee for Space Data Systems, 2012, pp. 1–1). Formats, format technologies, and their lifecycles have captured the attention of the digital preservation field since its beginnings (Rothenberg, 1999; Waters & Garrett, 1996). Much research in digital preservation has emphasized the notion of file format obsolescence and the perceived necessity of interventions (Lawrence, Kehoe, Rieger, Walters, & Kenney, 2000; Rothenberg, 1999; Ryan, 2014; Ryan, Graf, & Gordea, 2015; Waters & Garrett, 1996). However, surprisingly little is known about the actual patterns of evolution of core technologies such as file formats and their features.

Digital stewardship aims to decouple the lifecycle of digital objects from the lifecycle of their constituent artifacts. The aim is to manage the lifecycle of information assets based on their value and use, rather than be constrained by the systems that are used to store and manage them or the technologies used to encode and perform them. Digital objects, performed on a computing system, are the result of interacting elements such as: bitstreams; software code; dependency structures; networks; hardware; peripherals; and descriptive, structural, and administrative metadata. These elements undergo lifecycles of differing speeds according to factors such as market innovation and technological evolution. Understanding these factors is a prerequisite for effective actions. However, very little evidence exists that describes lifecycles and other phenomena associated with obsolescence and provides robust methods for assessing the state of any given format technology. The lack of evidence about how technology and its use evolve creates doubts in the methods that are proposed to mitigate the associated risks.

This article aims to present and evaluate a methodology that allows us to investigate how format technologies evolve in networked online markets. To do this, we address two main research questions: *(RQ1) How can longitudinal data*

extracted from the web provide indicators of format trends? (RQ2) Does the evolution of format technology usage follow identifiable patterns?

We propose a longitudinal analysis as a stepping stone for understanding and quantifying how technological lifecycles evolve. We are, at this stage, interested in identifying methods to describe the changing popularity of format technologies in identified markets. To this end, the following sections develop four major concepts.

1. We aim to identify a quantitative model of technology adoption applicable to the evolution of competing technologies in networked markets that can be populated with historical data to analyze the adoption of individual technologies. **Diffusion theory** and market research have developed models for analyzing the evolution of different products' share in a common market. In particular, the *Bass diffusion model* has proven to be robust and broadly applicable.
2. A conceptual model of key format technology components as a market system that maps format technology concepts to the theory of diffusion quantified in the Bass diffusion model enables our **Technology Lifecycle Analysis** to represent and analyze the evolution of components.
3. **Longitudinal datasets** must cover a minimum number of data points describing market elements. The historical data in web archives are a crucial source, but noisy and incomplete.
4. **Analysis** methods are needed that can be used to apply and test models on these datasets. Our workflow combines data cleansing, conflict resolution, non-linear regression, and leave-one-out cross-fold validation.

In the final sections, we evaluate the reliability of the method and the fit of the resulting models; compare results across markets; and discuss implications of the results.

Background

Obsolescence and technology diffusion

Digital preservation researchers have long focused on finding efficient methods to mitigate the effects of technology evolution. It is commonly accepted that obsolescence is a key risk (Vermaaten, Lavoie, & Caplan, 2012) and, correspondingly, it is one of the central factors in the environment of digital preservation that have to be monitored for potential change (Becker et al., 2012; Conway et al., 2012). This has resulted in initiatives ranging from simple format registries such as PRONOM to automated monitoring systems such as AONS (Curtis, Koerbin, Raftos, Berriman, & Hunter, 2007) and Scout (Becker, Faria, & Duretec, 2014). Some organizations have conducted limited experiments on their holdings to determine whether the objects are accessible (Holden, 2012). Pearson and Webb also state interest in "identifying file formats ... that can no longer be rendered (and are therefore obsolete); and those that are likely to become unrenderable within a timeframe demanding action (and are therefore obsolescent)" (2008, p. 93).

In economics, technology evolution has been described using concepts such as network effects and path dependence. Path dependence highlights the far-ranging effects that early events can have on later developments (Arthur, 1989, 1994). Where network effects exist, the value of a technology rises with the size of its network (Brynjolfsson & Kemerer, 1996). In this line, Rosenthal (2010) discusses how maturing markets and the web changed the market segment of desktop publishing. In the immature 1990's market, multiple competing formats existed, and obsolescence was common. However, as the dominant network's value rises, all vendors have an interest in compatibility with this network, and the dominant vendor often makes it easy for others to migrate to their technology. As a result, early events can cause a path dependent lock-in to a particular technology. Rosenthal argues that the formats that became obsolete were primarily those that did not gain sufficient early market share.

While Rosenthal asserts that "format obsolescence is a rare problem that happens infrequently to a minority of unpopular formats" (2010, p. 208), no solid evidence has been presented to either confirm or refute the existence of obsolescence as a real, as opposed to perceived, threat to digital assets. The argument describes a typical technology diffusion curve. Its focus is on 'widely used' formats within one specific, highly saturated market. How can we

characterize the evolving technologies in the long tail of data assets held in archives and repositories or in the emerging markets of mobile applications?

No robust general theories or models exist to describe, explain or predict obsolescence in digital ecosystems. However, obsolescence is not an inherent property of any particular object or technology, but of the relation between multiple components: An object can be considered obsolete *in a certain environment* if the environment does not possess the capability of accessing the object or, economically, if the costs of access outweigh the object's market value. Recent work has articulated the conceptual relationships between components such as formats, rendering software, and hardware dependencies, to support coordinated management and transition efforts in the face of each components' evolution (Dappert, Peyrard, Chou, & Delve, 2013; McKinney et al., 2014). However, the lifecycles of these components and their relationships have not been studied effectively. The absence of solid theories means that much well-intended research on risk assessment, decision support and preservation management is unable to provide effective measures of success and considerable efforts are spent without clear results.

The need for strengthening the evidence base is widely acknowledged. For example, the 2015 Agenda of the National Digital Stewardship Alliance (NDSA) highlights the urgency of establishing systematic evidence through publicly available datasets and repeatable processes (NDSA, 2014). In this spirit, the focus of this article is not to prove or disprove obsolescence, but to provide a key element of a theory of technology evolution by modelling lifecycles of technology *use*.

Diffusion theory

Technology diffusion theory is commonly traced back to Rogers, who in 1962 synthesized prior research in numerous studies into a theory of the diffusion of innovation that proved highly influential. In contrast to economists' attempts to develop explanatory theoretical models based on causal effects that are hard to verify, diffusion theory focuses on characterizing observable behaviours of adoption. Diffusion is defined as "the process by which an innovation is communicated through certain channels over time among the members of a social system" (Rogers, 2003, p. 11). Rogers distinguished four main elements: The *innovation* is an idea, in particular a technology, which is perceived as new. *Communication channels* are the means by which the new idea travels between members. *Time* measures the adoption rate of an innovation, with five suggested member categories based on adoption speed: innovators; early adopters; early majority; late majority; and laggards. Finally, the structure of the *social system* in which diffusion occurs affects the diffusion process in manifold ways (Rogers, 2003).

Diffusion theory was particularly influential in market research, where models were developed in order to analyze and predict product trends. The *Bass diffusion model* (Bass, 1969) introduced quantitative modelling and prediction and thus operationalized Rogers' diffusion theory. The basic assumption is that at any given time, the probability of an initial purchase of a product is dependent on the number of previous purchases made by others (Bass, 1969). Originally focused on consumer durables, the model has since been successfully applied to numerous other markets (Mahajan, Muller, & Bass, 1990) and was widely adopted for broader purposes of diffusion modelling (Ford, Menachemi, & Phillips, 2006; Mahajan et al., 1990; Van den Bulte, 2002; Wong, Yap, Turner, & Rexha, 2011).

As highlighted by Peres et al. (2010), diffusion is "driven by social influences' that include all of the interdependencies among consumers that affect various market players with or without their explicit knowledge" (p. 91). The Bass model focuses on how this is driven and distinguishes two types of potential adopters: innovators and imitators. While innovators adopt a certain product independently of previous adoption rates and based purely on external influences, the decisions of imitators are dependent on the number of previous adopters (internal influences). As the initial model was focused on durable products such as cars and refrigerators, product sales were a good indicator of overall adoption. The model in Eq. 1 presents the number of units sold at a specific point in time $S(t)$ as a function that is linearly dependent on previous adoption according to three parameters (p,q,m). The total number of ultimate adopters (m) denotes the *market potential*. The ratio between the *coefficient of external influence* (p) and the *coefficient of internal influence* (q) leads to identifiable adoption patterns.

$$S(t) = m\frac{(p+q)^2}{p}\frac{e^{-(p+q)t}}{(1+\frac{q}{p}e^{-(p+q)t})^2} \qquad (Eq.1)$$

Calculating p and q given a set of adoption rates allows us to predict the time to peak diffusion rate and the so-called 'tipping point' at which the adoption rate peaks.

The upper part of Figure 1 shows an example diffusion curve for a case in which internal influences are significantly higher than external influences. The lower part shows the change rate of new adoptions, which is the first derivative of Eq. 1. The resulting pattern shows a fast growth process followed by early decline, caused by the fact that the coefficient of internal influence is weak, so diffusion is not sustained.

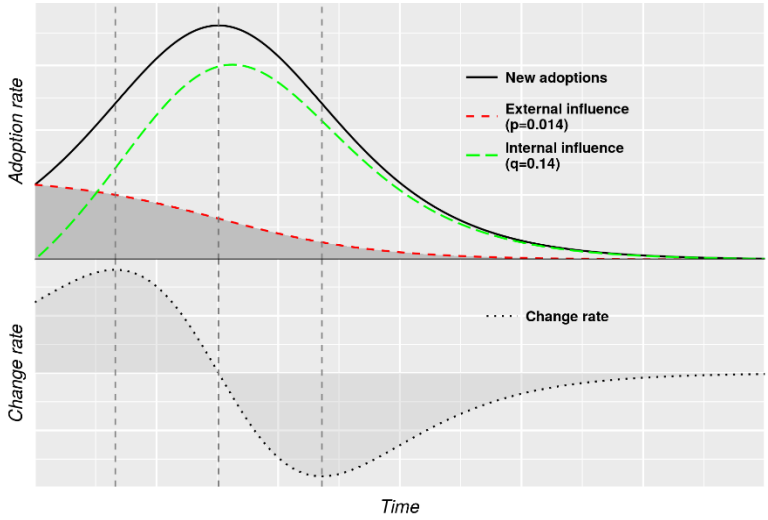

Figure 1 External and internal coefficients in the Bass model

Figure 2 shows four patterns of possible lifecycle curves distinguished by the ratio q/p between internal and external influence. The model of Figure 1 ($q/p=10$) is similar to the upper right. A different pattern with much slower, but sustained growth emerges with low external influences (e.g. a low number of innovators in the social system) and high internal influences (e.g. strong network effects), as shown below.

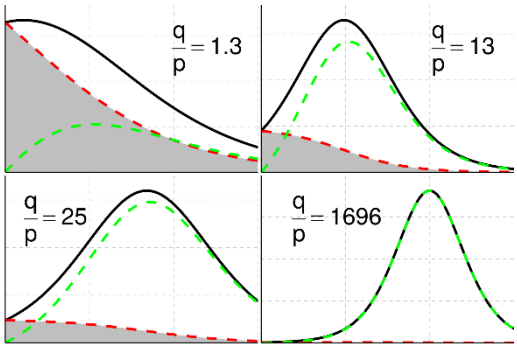

Figure 2 Patterns distinguished by the ratio between p and q

A minimum of three points is required for parameter estimation. Studies concerned about the reliability of estimates obtained with few data points suggest the use of nonlinear least squares (NLS) (Mahajan et al., 1990). The reliability of parameter estimation also depends on whether the peak is included in the data set, since the peak supports the determination of the market potential (Heeler & Hustad, 1980; Srinivasan & Mason, 1986).

If the lifecycle curves of format technologies can be described effectively with the Bass diffusion model, this opens up analysis and modelling techniques and allows predictions about when a format will no longer be actively used. Market diffusion models could thus provide means for understanding and predicting the evolution of different

products and answering questions such as: *Is this technology in the stage of growth, maturity, or decline? What is the median number of years that pass until a technology reaches its peak diffusion? Can we distinguish patterns of technology lifecycles? Do they correlate to particular constellations or events in technology development?* Shared evidence about these lifecycle patterns can eventually provide invaluable information for risk assessment and decision making.

Information Technology Evolution

In IT markets, the number of technologies and their relationships provide an analysis challenge. While some have developed methods for understanding technology evolution, there exists no systematic study of format lifecycles. Wong et al. (2011) use parameter estimation for the Bass model to analyze adoption patterns of internet-based communication tools such as email over a ten-year period. The data source is a survey with about 600 respondents. Adomavicius et al. (2008) develop an alternative approach to modelling evolution that uses dominant technology roles within a specific ecosystem to identify patterns of technology evolution without external factors such as market dynamics. The focus is on IT management decisions, not understanding technology lifecycles.

Longitudinal data sources

Substantial longitudinal datasets are the key ingredient for a quantitative lifecycle study. The usage of web archiving to understand historical trends is a new phenomenon (Milligan, 2012), but the size of web archives and the time span they cover makes them rich datasets for different kinds of socio-technical analysis, and the temporal analysis of their content is increasingly common (http://www.temporalweb.net/).

General conclusions based on web archives analysis need to consider coverage. Ainsworth et al. (2011) conducted an experiment and concluded that 35-90% of public URIs were archived at least once, and AlSum et al. (2014) showed that the Internet Archive has the broadest coverage with a recall value of 0.96.

The availability of characterization tools such as Apache Tika (http://tika.apache.org/) means that a rich set of metadata can potentially be extracted to enable quantitative analysis of the technical aspects of web archives holdings. Jackson provided a simple overview of different formats over time in the UK web (Jackson, 2012), and the data underlying this overview have been made available (The UK Web Archive, 2013).

While the currently available datasets are still of limited expressiveness, they provide sufficient detail to demonstrate the feasibility of the approach advocated in this paper. They cover a sufficient period of time in a relatively well-defined domain and provide, for each year, a histogram of the distribution of formats. However, web archives come with significant biases that make it challenging to derive reliable results.

Technology Lifecycle Analysis

Rogers' diffusion theory provided a qualitative description of an emerging understanding of the diffusion of ideas in a social system. Bass' quantitative model initially focused on the introduction of new classes of products.

Our hypothesis is that it should be feasible to apply the Bass diffusion model to the popularity of formats in a defined scope. To develop and test this argument, a clearer understanding of the relationships between format popularity and technology markets is required. We explain here the conceptual mapping of technology ecosystems to market research concepts. Table 1 clarifies key concepts, illustrates each with an example from traditional product sales, and maps the concepts to the domain of format technologies.

Concept	Definition	Example 1 (durable consumer products, typically one per customer)	Example 2 (format technologies, many per customer)
Innovation	A new idea or product	A new wireless router	A new file format (such as PDF 1.6)
Adoption	An individual's decision to adopt an innovation.	The initial purchase of a wireless router	The use of a format technology when creating a file
Release	The date an innovation is available for adoption	The date a product is available for purchase	The date a format specification and implementation is made available
Product lifetime	The time period between product release and product unavailability	Product released for sale on May 1, 2004; company stops offering it on April 30, 2007. Sales of remaining items continue. The last item is sold on April 1, 2010.	Format specification and reference implementation released on May 1, 2004; first file in this format released on June 1, 2004; company drops technical support April 30, 2007. Users continue to create files in this format until the last known file is created on April 1, 2016.
Likelihood of adoption of *P* at *T*	The probability that an Adoption of product *P* will be made at *T*	The probability that a consumer will purchase a wireless router at *T*	The probability that a file created at *T* will be represented in a specific format
Market segment	The scope of analysis is defined by type of product and region of interest. Overlapping segments can be defined.	wireless routers in Ontario; electronic consumer goods in North America; smartphones with 6" screen and Android 6.0 in South Korea; …	Electronic documents created in the UK; born-digital photographs in raw formats taken in Finland; CAD drawings created in Europe; Canadian web pages; …
Market potential	The total number of adopters within a Market Segment at the end of the period of interest	The total number of wireless routers in Ontario at the end of 2015	The total number of all electronic documents released in the UK until the end of 2015
Indicator	The indirect aggregate indicator used to determine actual adoption	Sales figures reported per time period per product	Relative frequency of product within file format profiles of web content per year

Table 1 Mapping technology diffusion concepts to format technology markets

In market research, an individual's *adoption* of an *innovation* typically translates into a purchase decision. In technology lifecycle analysis, it translates into an individual's choice of a format or feature to store an object.

An innovation can be adopted any time after its *release*, i.e. its availability for purchase or use. Just as products will still be available for purchase after production and sale is stopped by the producing company, formats can be used to create files after the originating organization drops technical support.

The *likelihood of adoption* refers to the probability that a consumer will buy the product. For formats, it denotes the probability that a digital object will be represented in a particular format. The *market potential* of a given product is the number of total items sold. For formats, it is the total number of files using a given technology. The *indicator* used to estimate the adoption rate of any particular format within a market segment is the percentage of files created that use this format as a share of all the files in the market segment during the specified period. Thus, timestamped format profiles in our analysis take the place of sales figures in marketing research (Bass, 1969).

Market segments are defined to scope the analysis and allow focused comparison. Segments are often hierarchically organized, but can be chosen by the analyst as long as they are clearly bounded. This is facilitated through fine-grained characterization of products and technologies. The technical properties of files, measured via characterization tools, range from basic identification data such as MIME type and format version to detailed characteristics such as the software module used to create the file. The hierarchical nature of the segments of interest is illustrated in Figure 3.

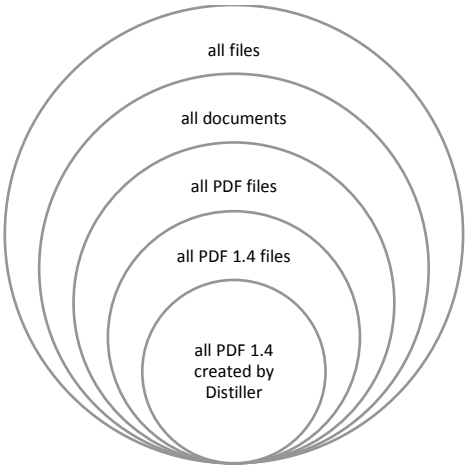

Figure 3 Hierarchical market segmentation (example)

We thus define market segments on three levels:

1. Segments of *content types* such as documents, images, and videos, where products are format families such as Adobe PDF.
2. Segments of *format families, where* products are individual versions of formats.
3. Segment of *tools for formats,* where products are software tools. We defined one such segment to evaluate whether the diffusion model can describe the lifecycle of such applications' use in creating particular formats, such as Adobe Distillers' share of PDF files.

Datasets

The dataset requirements for our proposed analysis can be divided into several categories:

1. **Open**: The dataset needs to be publicly available in order to support reproducibility.
2. **Significant time span**: To support trend analysis, the dataset needs to cover a time span longer than the duration of major lifecycle phases in the market. In the case of formats, the working hypothesis is that datasets with time spans of 10 years and more provide reasonable length to allow insights into the evolution of market diffusion. Our results confirm this assumption.
3. **Diverse**: The dataset should cover a diverse set of formats and their features. This is especially important as we are interested in understanding broad trends in different groups of formats.
4. **Representative**: The selected dataset should provide a useful characterization of the IT market of interest.

Longitudinal datasets that satisfy the above criteria are still scarce. However, web archives increasingly provide an effective source of data and currently are the only available datasets which contain diverse, heterogeneous data covering significant time periods of over a decade.

Even in cases when archives cannot provide full access for legal reasons, aggregate metadata about the overall technical statistics can be shared openly, as has been done by the UK Web Archive. The primary obstacle, then, is extracting the statistics from the archive holdings. This is increasingly feasible (Holden, 2012) and has been done for a substantial set of the UK Web Archive (Jackson, 2012). For example, a dataset from the Danish web covers 8 years in 441 million files (Becker, Faria, & Duretec, 2015). It contains in-depth, detail-rich feature sets extracted by the File Information Toolset (FITS), but covers fewer years. We use the UK Web Archive formats dataset describing 2.5 billion files (The UK Web Archive, 2013) due to its longitudinal coverage. For each of the years 1994-2010, a number of entries describe the number of items in each format encoded in a tab-separated file.

Two data quality issues need to be addressed: sampling irregularity and identification conflicts. Figure 4 shows the substantial variations in annual harvest sizes that cannot be explained by the growth of the UK web itself, but must be attributed to selection and crawling strategies. This means we cannot rely on unit numbers over the years. Instead, our adoption rate indicator will be computed relative to each year's volume.

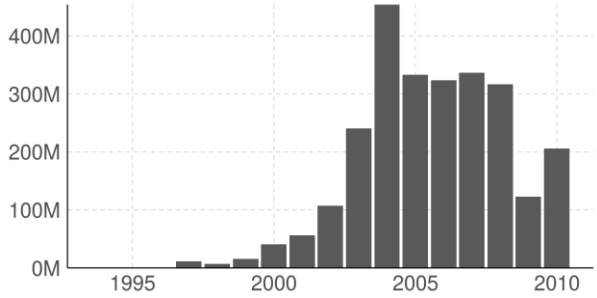

Figure 4 UK web Archive volume per year

The second issue arises because the MIME types reported by the server and the characterization tools frequently conflict. Table 2 provides three rows to illustrate the data structure. Each unique combination of all properties is counted separately. Of the 530,476 lines, 289,883 lines exhibit conflicts in the MIME types. Identification tools are not perfect (Holden, 2012) and will often deliver conflicting results (Kulmukhametov & Becker, 2014; Kumaravel, Dearborn, Witt, & Kuang, 2016). However, these conflicts occur in patterns according to a power law distribution, so we can identify frequently occurring conflicts that can be explained and resolve them using rule-based techniques (Kulmukhametov & Becker, 2014).

Server-reported MIME type	Tika-reported MIME teype	DROID-reported MIME type	Year	Count
application/ octet_stream	image/jpeg	image/jpeg; version=1.01	2006	11
application/pdf	application/pdf; version=1.2; software= "Acrobat Distiller 4.05 for Windows"	application/pdf; version=1.2	2007	5
text/plain	image/gif	image/gif; version=1989a	2004	2461

Table 2 UKWA format profile data examples

To calculate technology lifecycles, we also need a specification of markets that partitions the set into meaningful segments, such as the market of known image formats, and provides product release years to enable a unified mapping of multiple format statistics onto an age-adjusted curve. We compiled a spreadsheet that combines multiple formats into groups to define market segments of interest. All segments and products are available (Duretec & Becker, 2016b).

Analysis method

The following describes a systematic approach to leveraging longitudinal web archive data to support lifecycle analysis of format technologies on the web. The aim is to provide a stepping stone toward shared evidence and theory-building around technology evolution. To this end, the method is fully disclosed; all input and intermediate data and results are shared according to the FAIR principles (https://www.force11.org/group/fairgroup/fairprinciples); and all code, implemented in R, is made available (https://github.com/datascience/FormatAnalysis).

The following describes the five phases of the workflow labeled across the top of Figure 5: 1) Input preparation and configuration, 2) Preprocessing, 3) Parameter estimation, 4) Evaluation and model selection, and 5) Results and validation.

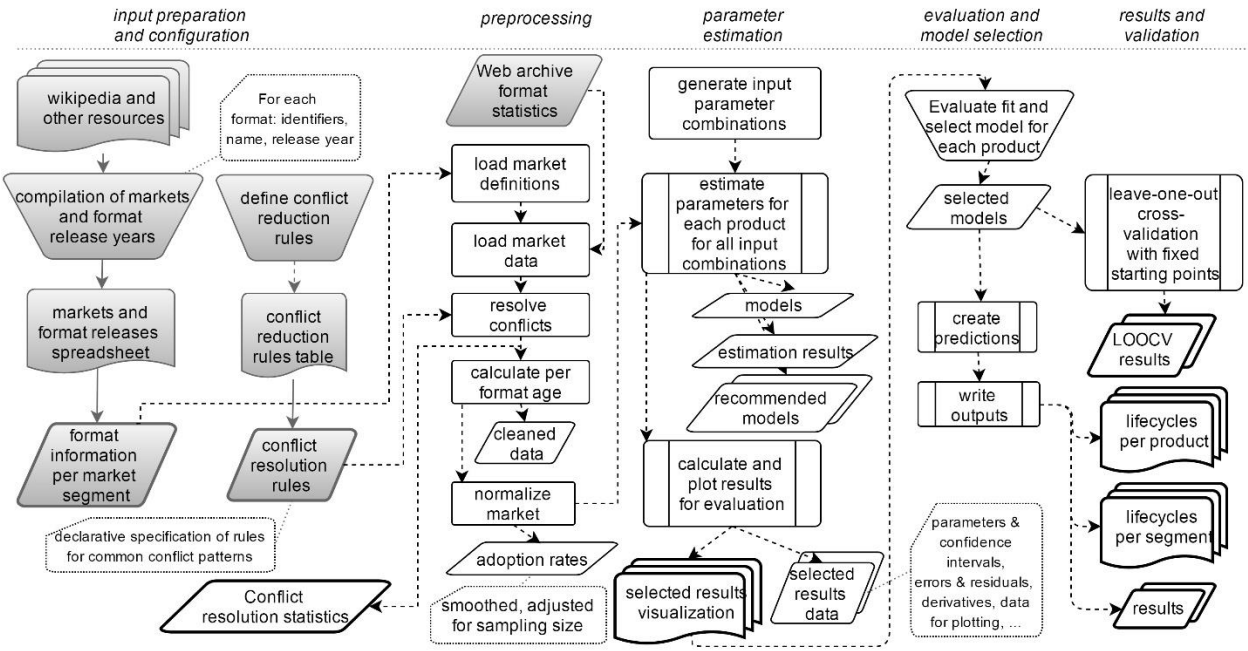

Figure 5 The analysis workflow comprises preparation (grey), processing and analysis, and validation

Input preparation and configuration

The workflow begins with the preparation of two sets of input data. These are compiled in an online spreadsheet (Duretec & Becker, 2016a).

Market segments are defined by sets of values for properties such as *MIME type* and *version*. Each sheet defines one market segment, and each line uses a combination of values to define one product within the segment. This structure allows the flexible combination of properties in varying levels of granularity to analyze different products such as the PDF format (the shares of all PDF within all documents), or one PDF version within PDF, or software products (the share of Distiller in producing PDF documents). For each product, a release year can be defined. Products with release years undergo lifecycle analysis. The segment size is determined by the aggregate of all listed products.

Conflict resolution uses declarative rules to resolve known conflicts in the dataset. This is essential for subsequent analysis since some segments contain large shares of conflicted items. The implemented mechanisms go beyond the data cleansing performed by Jackson (2012) and build on earlier work (Kulmukhametov & Becker, 2014). An analysis of conflict patterns supported the declarative formulation of a relatively small rule set in the market spreadsheet so that conflict resolution is extensible and fully integrated. A format is 'identified' by a tool if the value returned is different from *null*, an empty string, or *application/octet-stream.* Rules are applied in the following sequence:

1. *Custom*: If the set of values match a custom rule, apply it.
2. *Clean*: If all three values are equal, return the value.
3. *Agreement*: If Tika and Droid agree on a format, return that value.
4. *Single*: If only one source managed to return a value, return that value.
5. *Conflict:* If the conflict remains unresolved, "NA" is returned.

Preprocessing

The preprocessing steps load market definitions, conflict rules and the dataset. As each product can include multiple labels, the first step merges these, followed by a conflict reduction mechanism using the declarative specification described above. The indicator for product adoption is calculated relative to the market segment for each year. The release years are used to normalize product adoption curves according to age. A 3-year moving average is used to reduce the effect of outliers and sampling irregularities.

Parameter estimation

We conduct non-linear regression applying NLS using the Levenberg-Marquardt algorithm as implemented in *minpack.lm (*Elzhov, Mullen, Spiess, & Bolker, 2015). To ensure a comprehensive search space, exhaustive starting points for parameter combinations are generated. The estimation computes a set of models, each a set of resulting parameters *p,q,m* with associated residual error and starting points.

Multiple candidate models can result at this stage, and a purely error-based selection of the best fit is inadequate. To recommend selected models for manual evaluation, the models are sorted according to residual error using RMSE. Five models are selected by traversing down from the best-fit model and adding models where at least one of the parameters *p,q,m* is at least 10% different. For each, the evaluation of the goodness of fit is supported by calculating error, *q/p* ratio and confidence intervals for parameters, and plotting confidence bands and residual error. The resulting visualization supports the effective evaluation by combining all information required to evaluate model fit. For each set of candidate models, an overview visual is generated.

Evaluation and Model Selection

The goodness of fit of the regression method is evaluated numerically and visually using the following criteria:

1. **Sensible parameters** include positive *coefficients p,q* and *market potential* m, including the lower bounds of the confidence intervals and extreme *q/p* ratios.
2. We consider the relative **magnitude of the confidence interval** for parameters.
3. We consider the **RMSE** of residuals.
4. The overall **closeness of fit** is evaluated visually.
5. Places where sampling noise appears to cause **jumps in the data points** reduce confidence in source data quality or market segmentation. (This occurred with some formats for which very little data existed).
6. The **number of data points** is considered. Formally, a minimum of 3 points is needed for fitting. At least 4 points are needed to calculate confidence intervals. In practice, more than 4 are often needed for reliable estimates.
7. Considering that diffusion needs time, a **peak** in year 0 is generally less realistic than a later peak.
8. We consider whether the **peak adoption point** is included in the data, since this influences estimation reliability.
9. Finally, the randomness of **residuals** is evaluated using a residual plot.
10. Throughout the process, we consider any possible **effects of smoothing**.

Results and Validation

Selected models are specified as an input file to the final calculation steps, which perform a leave-one-out cross-validation using the starting points that led to the selected parameters; create predictions for future years; and store the combined outputs and additional plots.

Application and Results

This section discusses the application of the analysis method to the UK web archive format dataset to analyze selected format technologies. All detailed intermediate and final results are available (Duretec & Becker, 2016b).

Market segments and conflict resolution

Table 3 shows two of the 59 custom rules created for this data set.

ID	Property	server	tika	droid	resolveTo
5	mime	application/msword	application/octet-stream	application/msword	application/msword
37	mime	image/pjpeg	image/jpeg	application/octet-stream	image/jpeg

Table 3 Custom conflict resolution rules

Table 4 illustrates how these and other rules are used to resolve example conflicts.

Server output	Apache Tika	Apache Droid	affected files	Resolved To	Rule Used
Application/ msword	application/ octet-stream	application/msword	2831765	application/ msword	Custom rule #5
text/rtf	application/ rtf	application/rtf	91659	application/ rtf	Agreement between identification tools
image/pjpeg	image/jpeg	application/octet-stream	71705	image/jpeg	Custom rule #37
image/gif	image/jpeg	image/x-pict	9277	NA	Conflict unresolved

Table 4 Exemplary conflicts in the data set and their resolution

We defined five market segments for formats, five for versions of formats, and one for tools. Table 5 summarizes the defined segments and the effect of conflict resolution on data quality.

Market segment	Number of products	Size (% of entire archive)	Size (%) after conflict resolution	Conflicted files (% of segment)	Number of custom rules	Remaining conflicts (% of segment)
ARCHIVE	7	0.062	0.058	61.186	0	6.334
AUDIO	32	0.061	0.057	49.036	12	3.820
DOCUMENTS	19	0.988	0.958	22.931	10	2.789
IMAGE	48	10.563	10.529	7.922	9	0.299
VIDEO	22	0.028	0.021	69.526	18	6.927
PDF versions	11	0.768	0.768	100	3	0.014

BMP versions	6	0.012	0.012	100	0	0.094
GIF versions	2	4.014	4.014	100	0	0
FLASH versions	7	0.071	0.071	100	0	0
HTML & XHTML versions	6	40.621	39.224	100	4	3.183
PDF generators	9	0.822	0.795	10.749	2	2.568

Table 5 Market segments and the effect of conflict reduction on data quality

We adopted a very defensive approach to rule definition to avoid distorting errors, so numerous conflicting patterns remain for which no resolution is specified. Still, similar to the findings in Kulmukhametov & Becker (2014), a small number of rules is highly effective in improving data quality. The effect on the ratio of conflicted values shows that residual conflicts are small. While some distortions can be expected for low-volume products, these should bear no relevance on the general findings.

Evaluation of fit

Estimation can yield multiple parameter sets. Where the same model is returned for all combinations of starting points, its fit is generally very robust. Figure 6 shows an example visualization generated to support judgment according to the criteria outlined above. It shows an almost ideal case, PDF 1.2 in the PDF market. Note that the points plotted in the main plot are raw data points, while the parameter estimation uses a moving average. This explains the outlier after the peak. For the adoption rate (y-axis), 1000 is equivalent to 100% market share. The range varies across diagrams. The area under the curve is proportional to the market potential reached by the technology.

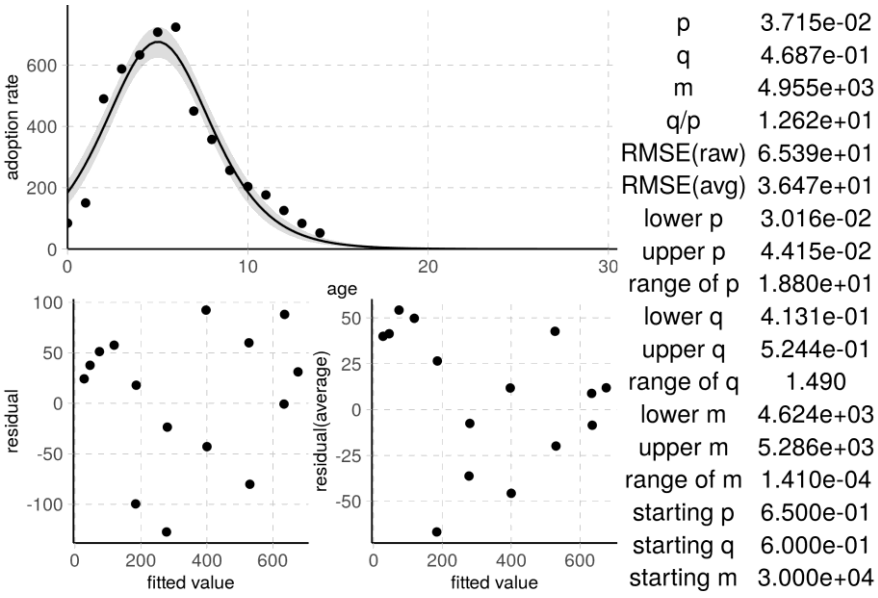

Figure 6 An excellent fit is found to describe PDF 1.2 (in PDF)

In this case, the presence of a full set of data points owing to the format's age leads to an excellent fit with robust parameters, low error and random residuals against initial data points (left) and moving-average points (right). The Bass model provides an accurate description, with *p=0.04* and *q=0.47*.

Figure 7 shows a similarly robust fit for HTML 4.01. Similar to others, it passes its peak around year 7, and its curve is well characterized in the Bass model (*p=0.01, q=0.55*).

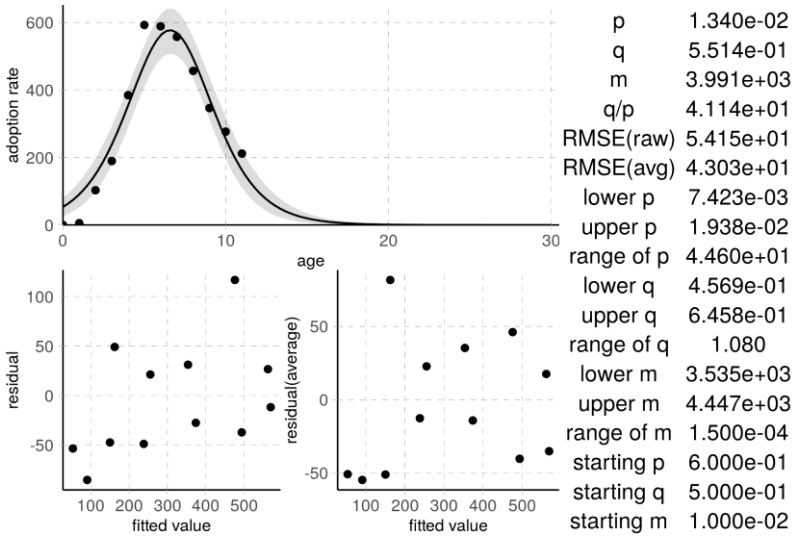

Figure 7 The best model found for HTML 4.01 (in HTML)

For other products, multiple candidate models are returned with varying degrees of fit. For example, Figure 8 visualizes the five Bass curves returned for the Graphics Interchange Format in the market of images.

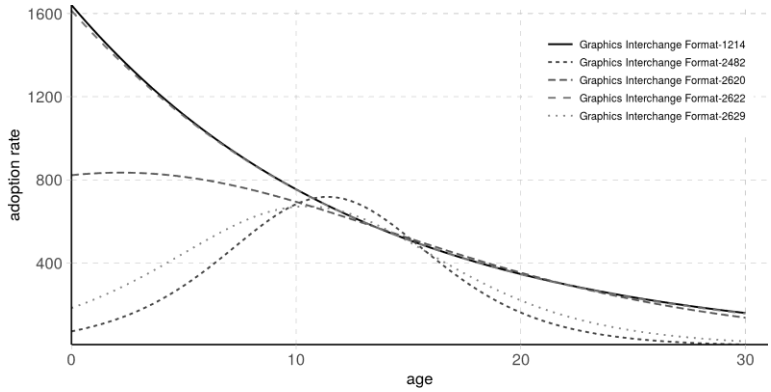

Figure 8 Candidate models for GIFs share within images

Using visualizations as above, two models are quickly dismissed based on plausibility, fit and error. Figure 9 shows model 2620 (*p=0.05, q=0.06*), which is selected and considered sufficiently robust despite the lower bound for q. It outperforms the remaining two models in several criteria such as *sensible parameters*, *R*MSE and residuals (lower left). Figure 10 shows the closest competing model (*p=0.02, q=0.22*), which assumes a later peak. For GIF, the early timing of the peak seems realistic.

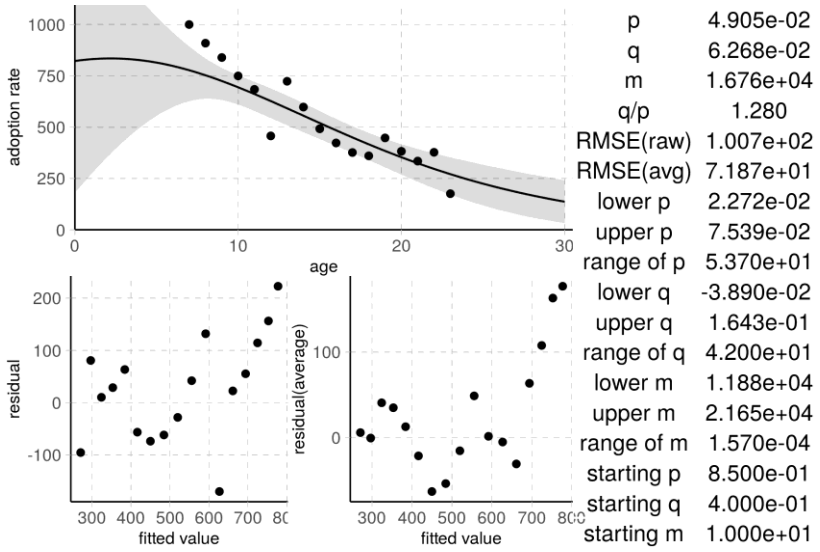

Figure 9 Evaluation of fit for model 2620 (GIF)

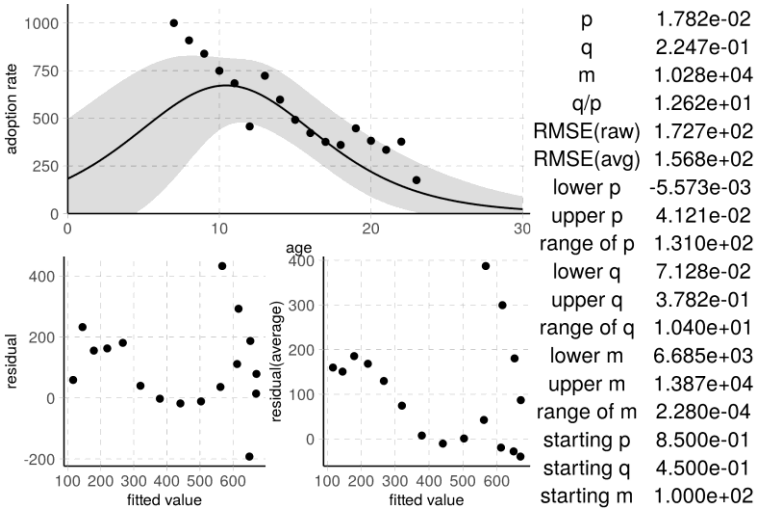

Figure 10 Evaluation of fit for model 2629 (GIF)

Cross-fold validation

To evaluate the generalization performance of our estimation method, we perform Leave-One-Out Cross Validation with the parameter starting points fixed to the selected models. To evaluate the results, we aggregate the ratio of points where the real value lies within the prediction interval. The aggregate percentage of successfully predicted points forms the basis of evaluation. The rate varies across markets, with most markets resulting in success rates between 0.85 and 1.0. As shown in Figure 11, series with 6 or more data points provide robust estimates. The size of the points denotes the number of products in that category.

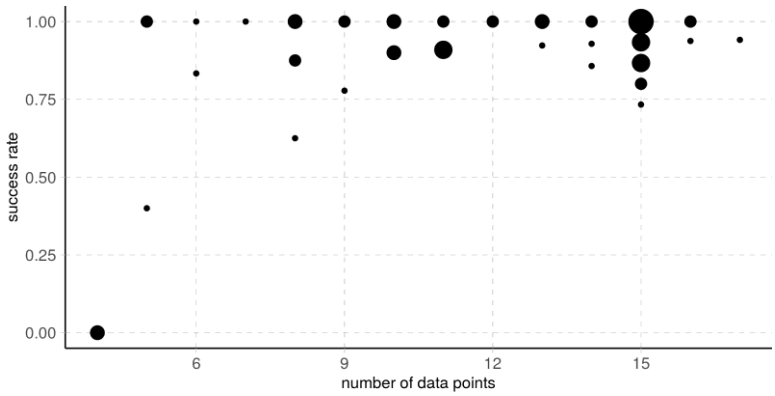

Figure 11 Cross-fold validation results indicate robust estimates

Limitations and conditions

Analysis failed to produce a reliable result in several cases, including PDF 1.7 and PDF A/1a within the PDF market. However, analysis only failed in cases where only 3 or 4 data points were available, or when neither the peak nor the inflection point were covered. The evaluation process identified these cases clearly.

Since the curves represent the lifecycles of the analyzed technologies on the UK web, they cannot by themselves be taken as representative of the overall use of these technologies in other markets, and it remains to be seen how similar the curves are for formats that are not typically used to publish content online. However, the analysis suggests that diffusion of such technologies follows a Bass model. Application of the approach to data outside the web domain could thus use the same method to analyze the holdings of legal deposit or institutional repositories, or federated repository statistics. This would provide valuable comparisons, shed light on opportunities for generalizing these findings, and enable informed estimates about patterns to expect in cases where insufficient data points prevent parameter estimation.

Patterns in the evolution of format technology usage

The results demonstrate that longitudinal data extracted from the web can provide reliable indicators of format trends that can be used to compute diffusion lifecycles, thus answering RQ1. This enables us to represent format technology evolution conceptually and analyze the resulting lifecycle curves for patterns. To address RQ2, we explore markets of formats, format versions, and tools. We focus first on the largest and frequently discussed market of documents and compare the two most popular formats in our list to two formats that lost most of their market share. We then explore differences between formats, versions and tools to evaluate if the method is applicable across these types. All presented models have passed the evaluation discussed above. Full results and figures for all market segments listed in Table 5 are openly shared as a data set (Duretec & Becker, 2016b).

Lifecycle curves for popular and unpopular document formats

When we consider cumulative market share over time, the most frequently occurring document formats in the data set are Adobe PDF and Microsoft Word. As the corresponding lifecycles in Figure 12 and Figure 13 show, their share in the document market has levelled off or dropped.

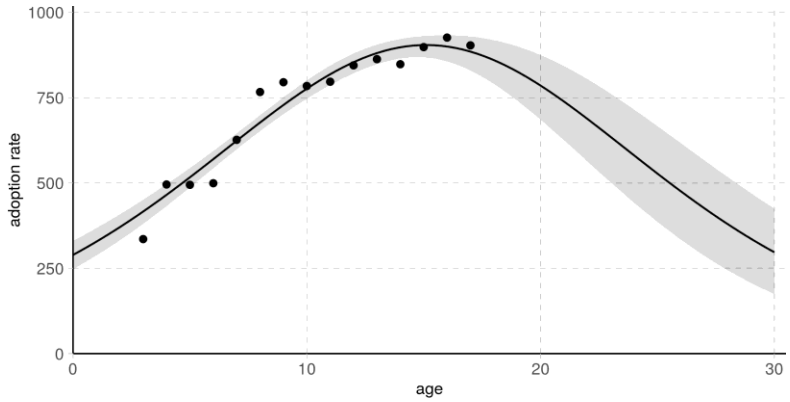

Figure 12 PDF is still strong, but has levelled off (83.6% overall market share).

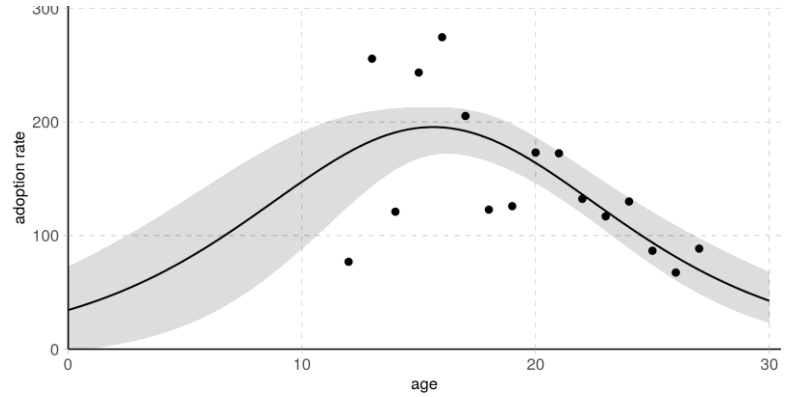

Figure 13 The binary MS Word format reached its peak at age 16 (13.7% overall market share)

These two formats came to substitute earlier formats designed for similar purposes, whose shares dropped correspondingly. Figure 14 shows that the lifecycle of WordPerfect, which pre-dates the web, all but disappeared after a low peak at age 20. Figure 15 follows the disappearance of the once dominant PostScript format, which lost over 91% of its former share between 2005 and 2010.

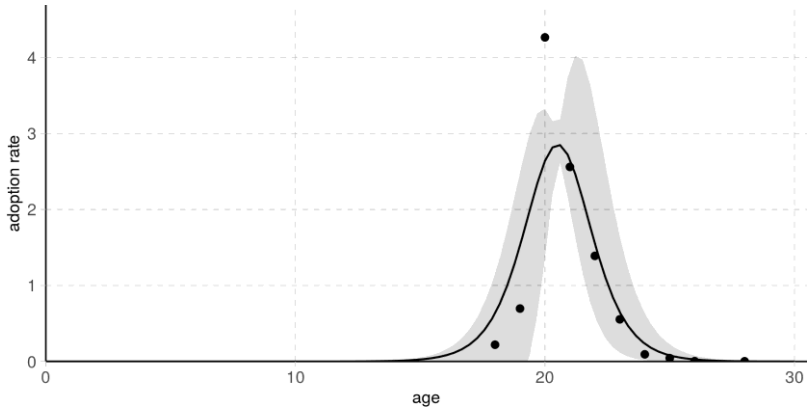

Figure 14 WordPerfect lifecycle (0.005% overall market share).

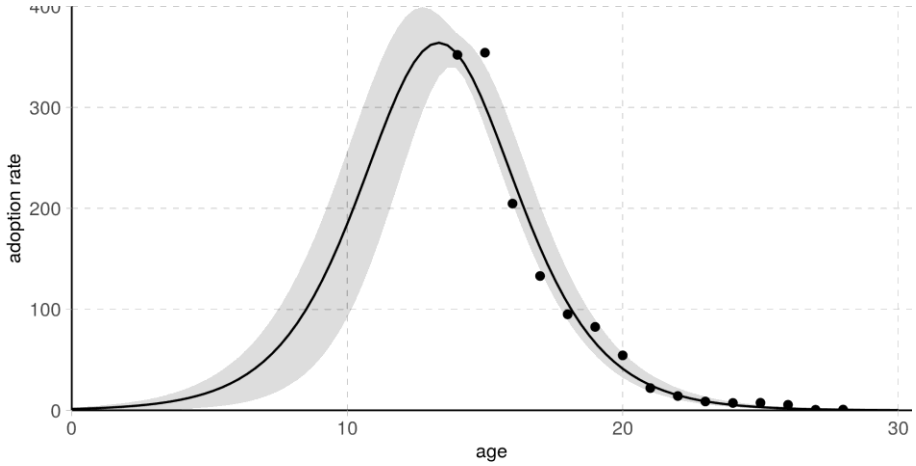

Figure 15 Adobe PostScript lifecycle (1.3% overall market share).

The life spans of formats, format versions and tools

Normalizing lifecycle curves according to age allows us to compare the length of different technologies' lifetime. As shown above, significant time passes before format groups such as PDF, whose versions are often backward-compatible, reach their peak and decline. In contrast, Figure 16 compares the versions of PDF where sufficient data allowed an estimate. Independently of their market potential, usage of most versions declined after 5 to 7 years.

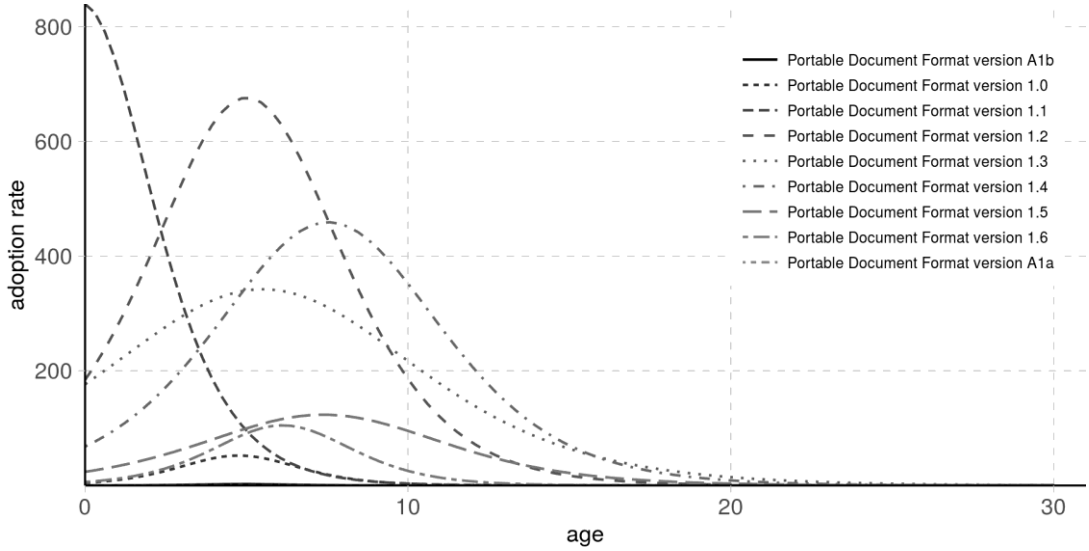

Figure 16 Diffusion speed of PDF versions in comparison

Figure 17 shows how successive versions of the Adobe Distiller, used for creating PDF files, have partially replaced their predecessors, while the initial version never became popular.

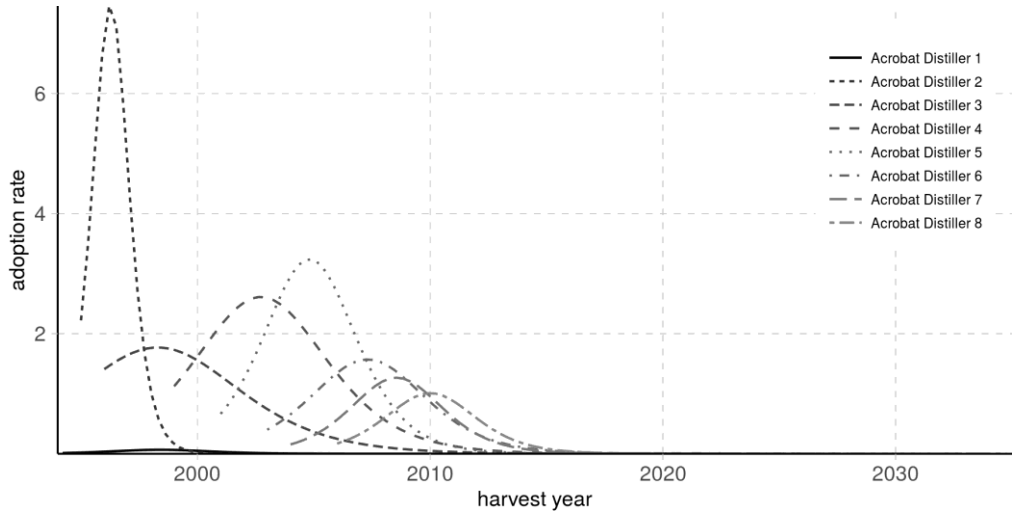

Figure 17 Successive versions of Distiller over time

Figure 18 aggregates these results across all market segments and plots expected life spans (represented by the time to peak) against product release date and market potential. It shows wide differences in life spans, with time-to-peak ranging from 3 to over 20 years for formats. Lifespans are not correlated with peak adoption rates, but for formats and their versions, are correlated (*r=-0.6*) to release dates (i.e newer formats reach their peak sooner). Note that for content not typically published on the web, these trends may look quite different.

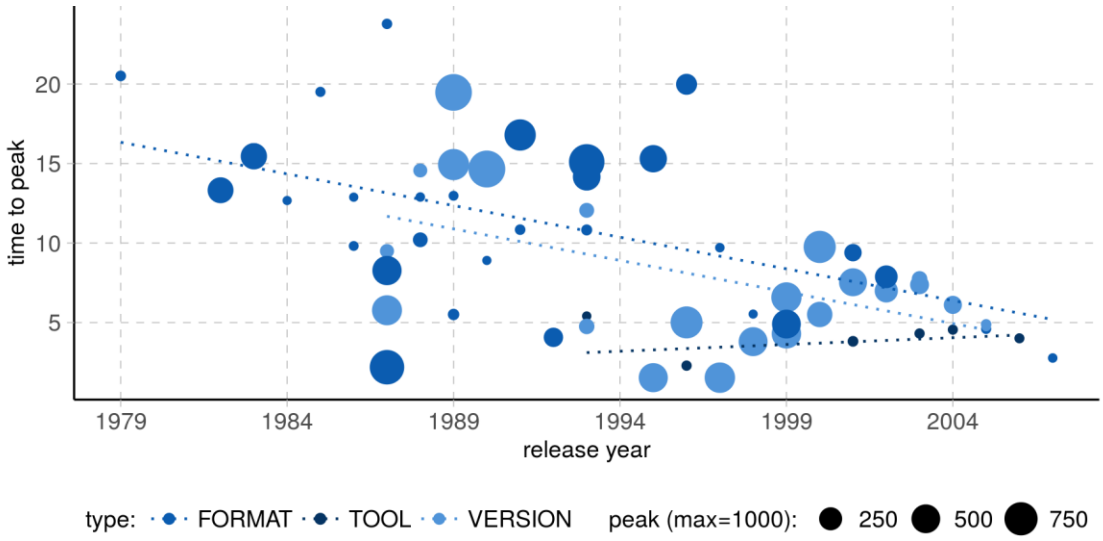

Figure 18 Life spans over time across markets

External and internal influences on diffusion

The representation of lifecycles in the Bass model allows us to establish and compare ranges for internal and external influences. For consumer durables, the coefficients typically average about (*p=0.03, q=0.38)* (Mahajan et al., 1990; Sultan, Farley, & Lehmann, 1990). To assess whether we can make similar generalizations, we analyze p and q for different groups. The average values for coefficients are (*p=0.03, q=0.59*) when all accepted models are considered, and (*p=0.05, q=0.5*) for excellent fits. The coefficients for all models are plotted in Figure 19 with a convex hull to highlight regions.

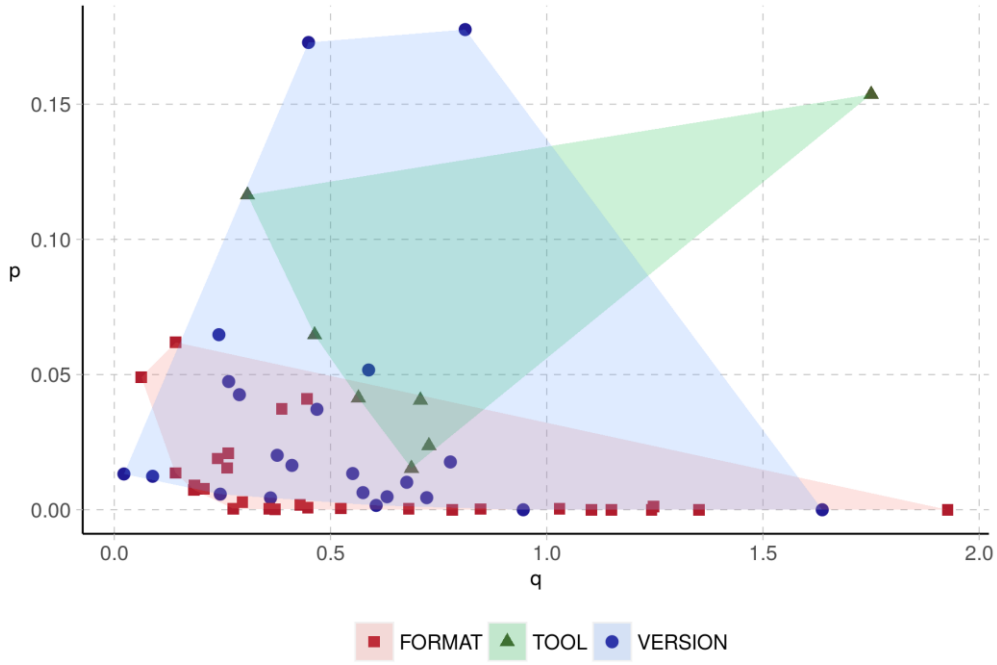

Figure 19 Regions for p and q, including outliers

The plot shows a significant difference between individual format versions, formats, and applications. However, the coefficient of internal innovation p is almost always within (0,0.1), with a few outliers up to 0.35. For groups of competing formats, p is always below 0.1. On the other hand, the coefficient of external innovation q varies substantially within (0.1,1.9). This can be expected considering adoption decisions are repeated so that multiple dependent adoptions are counted separately.

We expect that products with high market potential exhibit different patterns than those with low potential. Figure 20 indeed shows that products with low adoption are separated from those with medium and high market potential. The spread of q is largest with widely adopted products.

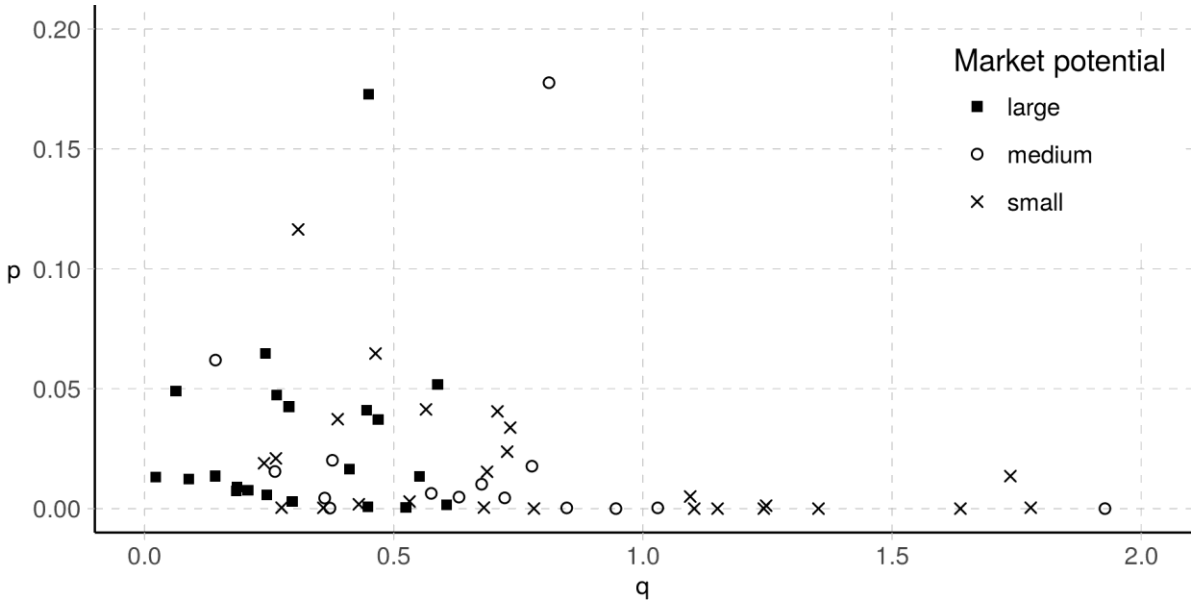

Figure 20 p/q clusters arise when distinguishing market potential

These figures suggest that lifecycle patterns differ according to market potential and product type. The *q/p* ratio between external and internal influence describes the curve shape. In fact, the q/p ratio of PDF (10) corresponds to Figure 1 (in the Background), while the ratios of GIF, TIFF, JPEG and PNG in the image market correspond to the ratios pictured in ascending order in Figure 2. When subsequent versions are introduced into closely defined markets such as HTML, PDF, or FLASH, the ratio increases, as illustrated in Figure 21. This is consistent with network effects.

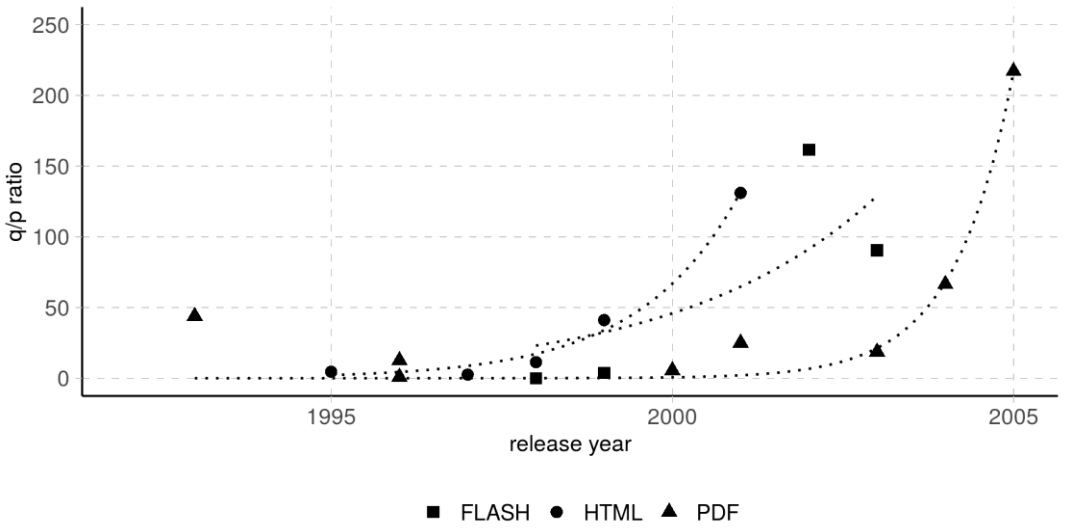

Figure 21 Trends in the q/p ratio for subsequent releases

Discussion

The results demonstrate that the Bass model is applicable to the evolution of format technologies in networked markets, and datasets from web archives can be used for parameter estimation. Non-linear regression as applied here is effective in identifying the fit of models to empirical data, although it requires careful numerical and visual judgment. The resulting models describe format evolution trends remarkably precisely.

The number of years until a format reaches the end of its lifetime varies considerably. By calculating its point on the Bass diffusion model and considering the change rate of the current period, we can identify whether a particular format is in a stage of growth, maturity, or decline.

The analysis identifies that the format technology lifecycles we studied indeed follow identifiable patterns that differ according to at least three characteristics: (1) Technologies with high market potential differ from those with low potential. (2) Formats differ from format versions and tools. (3) Subsequent versions differ from initial releases.

The quality of estimates obtained with fewer data points and the longitude of lifecycles shows that the time period we used is more than sufficient to conduct a meaningful analysis of this kind. All series with 6 or more points performed equally in terms of RMSE and cross-fold validation. This suggests opportunities to extend this analysis to data sets covering shorter time periods.

The initial narrowing of parameter boundaries should be corroborated with different data sources. Comparisons can be made across any market segment and cluster by adapting the definitions in the spreadsheets. Further analysis can make more fine-grained distinctions between lifecycle patterns. What is notable is that the Bass diffusion model allows us to express these processes and leverage the underlying theory to characterize and explain technology evolution.

On a practical level, a better understanding of lifecycle patterns can contribute to forecasting and planning for repository management. Caution is advised: A lack of popularity does not equal obsolescence, as rendering is often supported beyond a technology's active lifetime. However, the lifecycles provide a stepping stone to understand obsolescence. The patterns could be used to forecast expected growth and develop ingest capabilities in time, and to identify the optimal time for interventions. For example, identifying at which lifecycle stage tool support is typically best would make it easier for repository managers to schedule format validation and preservation planning tasks. We speculate that this point lies at the second inflection point, when the majority of adoptions have taken place and the change rate reaches its minimum (see Figure 1).

Since this article is the first of its kind, many opportunities exist to improve on and extend our current methods.

First, analyses on additional datasets should cover additional market segments and products and investigate whether the findings from the UK web can be corroborated.

- The current set of market segments should be extended. For example, video formats are generally considered containers. Evolution and incompatibility take place on the level of codecs, which have not been covered.
- Since the timespan conditions for longitudinal datasets suggest that 8 years coverage should enable robust analysis, data sets such as the feature-rich Danish web dataset can be used, and the method could be applied to (federated) repository statistics.
- Advanced conflict reduction rules are needed to improve data quality to a degree that supports more reliable parameter estimation for low-volume technologies most prone to obsolescence.

Second, future research could leverage extensions of the Bass model that integrate concerns beyond innovators and imitators to develop robust explanations of the factors contributing to growth and decline. This can include flexible market potential and the variable nature of internal influence (Mahajan et al., 1990) and consider alternative models.

Third, the lifecycles of multiple components should be studied in conjunction. Format lifecycles should be correlated to the diffusion of format support in the social system of vendors and to important events to evaluate their impact. For example, Flash and Java applets are becoming obsolete (i.e. the rendering costs are increasingly higher than access value for many users). By placing significant events such as the end of support of leading platforms on the diffusion lifecycle, we may evaluate how these events influence popularity. Similarly, the occurrence of rendering problems (Cochrane, 2012) could be compared to lifecycle stages. Comparison of findings can identify patterns, analyze how these change across time, and consider more fine-grained feature sets. Detailed analysis can focus on individual products of interest and describe how they become obsolete.

Research on predictive models could explore to what degree popularity can be used to predict obsolescence. We have included prediction in our method and the results data set, but regard these as initial steps towards an explanatory and predictive theory that supports forecasting of trends and impending technological change.

Conclusions

This article presented a reproducible and extensible analysis method that provides a robust baseline for computing format technology usage lifecycles. Our findings show that format characterization tools can extract usable aggregate longitudinal data from web archives that provide effective indicators of format trends. Web archives thus constitute invaluable historical data sources that cover sufficient time periods to conduct robust analysis. The Bass model provides an effective descriptive model for how popularity of format technologies evolves over time, and the resulting models produce reliable predictions. We evaluated the reliability, implications and limitations of the method, and elicited conditions for the method to be applied.

The evolution of format technology follows identifiable patterns that can be characterized using diffusion models. The method allows us to move from visualizing yearly trends towards explicitly modelling and representing change

in format technology according to diffusion theory. The curves can be used to forecast ingest trends and future needs for repositories and archives who receive digital content with a delay.

This demonstrates the feasibility of a data science approach to understanding technology lifecycle evolution, provides a baseline for more advanced analytics techniques and methods, and opens up opportunities for diverse analyses of similar kinds on other data sources and additional technologies.

Acknowledgements

Part of this work was supported by WWTF through BenchmarkDP (ICT12-046) and by NSERC through RGPIN-2016-06640.

References

Adomavicius, G., Bockstedt, J. C., Gupta, A., & Kauffman, R. J. (2008). Making Sense of Technology Trends in the Information Technology Landscape: A Design Science Approach. *MIS Q.*, *32*(4), 779–809.

Ainsworth, S. G., Alsum, A., SalahEldeen, H., Weigle, M. C., & Nelson, M. L. (2011). How Much of the Web is Archived? In *Proc. ACM/IEEE JCDL* (pp. 133–136). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1998076.1998100

AlSum, A., Weigle, M. C., Nelson, M. L., & Sompel, H. V. de. (2014). Profiling web archive coverage for top-level domain and content language. *International Journal on Digital Libraries*, *14*(3–4), 149–166. https://doi.org/10.1007/s00799-014-0118-y

Arthur, W. B. (1989). Competing technologies, increasing returns, and lock-in by historical events. *The Economic Journal*, 116–131.

Arthur, W. B. (1994). *Increasing Returns and Path Dependence in the Economy*. Ann Arbor: University of Michigan Press.

Bass, F. M. (1969). A New Product Growth for Model Consumer Durables. *Management Science*, *15*(5), 215–227. https://doi.org/10.1287/mnsc.15.5.215

Becker, C., Duretec, K., Petrov, P., Faria, L., Ferreira, M., & Ramalho, J. C. (2012). Preservation watch: What to monitor and how. *Proc. International Conference on Digital Preservation (IPRES)*.

Becker, C., Faria, L., & Duretec, K. (2014). Scalable decision support for digital preservation. *OCLC Systems & Services: International Digital Library Perspectives*, *30*(4), 249–284. https://doi.org/10.1108/OCLC-06-2014-0025

Becker, C., Faria, L., & Duretec, K. (2015). Scalable decision support for digital preservation: an assessment. *OCLC Systems & Services: International Digital Library Perspectives*, *31*(1), 11–34. https://doi.org/10.1108/OCLC-06-2014-0026

Brynjolfsson, E., & Kemerer, C. F. (1996). Network externalities in microcomputer software: An econometric analysis of the spreadsheet market. *Management Science*, *42*(12), 1627–1647.

Cochrane, E. (2012). *Rendering Matters-Report on the results of research into digital object rendering*. Archives New Zealand.

Consultative Committee for Space Data Systems. (2012). Reference Model for an Open Archival Information System (OAIS). CCSDS.

Conway, E., Matthews, B., Giaretta, D., Lambert, S., Wilson, M., & Draper, N. (2012). Managing risks in the preservation of research data with preservation networks. *International Journal of Digital Curation*, *7*(1).

Curtis, J., Koerbin, P., Raftos, P., Berriman, D., & Hunter, J. (2007). AONS-An Obsolescence Detection and Notification Service for Web Archives and Digital Repositories. *New Rev. Hypermedia Multimedia*, *13*(1), 39–53. https://doi.org/10.1080/13614560701423711

Dappert, A., Peyrard, S., Chou, C. C. H., & Delve, J. (2013). Describing and Preserving Digital Object Environments. *New Review of Information Networking*, *18*(2), 106–173. https://doi.org/10.1080/13614576.2013.842494

Duretec, K., & Becker, C. (2016a). Format markets. Retrieved from http://purl.org/dp/formats/lifecycle-analysis/marketsegments

Duretec, K., & Becker, C. (2016b). *Format technology lifecycle analysis based on the UK Web Archive Format Profile dataset*. Retrieved from https://dx.doi.org/10.6084/m9.figshare.c.3258991

Ford, E. W., Menachemi, N., & Phillips, M. T. (2006). Predicting the Adoption of Electronic Health Records by Physicians: When Will Health Care be Paperless? *Journal of the American Medical Informatics Association : JAMIA*, *13*(1), 106–112. https://doi.org/10.1197/jamia.M1913

Heeler, R. M., & Hustad, T. P. (1980). Problems in Predicting New Product Growth for Consumer Durables. *Management Science*, *26*(10), 1007.

Holden, M. (2012). Preserving the Web Archive for Future Generations. In *The Memory of the World in the Digital age: Digitization and Preservation*. Vancouver: United Nations Educational, Scientific and Cultural Organization.

Jackson, A. N. (2012). Formats over Time: Exploring UK Web History. *arXiv:1210.1714 [Cs]*. Retrieved from http://arxiv.org/abs/1210.1714

Kulmukhametov, A., & Becker, C. (2014). Content Profiling for Preservation: Improving Scale, Depth and Quality. In K. Tuamsuk, A. Jatowt, & E. Rasmussen (Eds.), *The Emergence of Digital Libraries – Research and Practices* (pp. 1–11). Springer International Publishing. Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-12823-8_1

Kumaravel, H. V., Dearborn, C., Witt, M., & Kuang, Y. (2016). Measuring the Accuracy of File Format Identification Tools. Presented at the 11th International Digital Curation Conference, Amsterdam: DCC.

Lawrence, G. W., Kehoe, W. R., Rieger, O. Y., Walters, W. H., & Kenney, A. R. (2000). *Risk Management of Digital Information: A File Format Investigation* (Vol. 93). Council on Library and Information Resources.

Mahajan, V., Muller, E., & Bass, F. M. (1990). New Product Diffusion Models in Marketing: A Review and Directions for Research. *Journal of Marketing*, *54*(1), 1–26. https://doi.org/10.2307/1252170

McKinney, P., Knight, S., Gattuso, J., Pearson, D., Coufal, L., Anderson, D., … Hutař, J. (2014). Reimagining the Format Model: Introducing the Work of the NSLA Digital Preservation Technical Registry. *New Review of Information Networking*, *19*(2), 96–123. https://doi.org/10.1080/13614576.2014.972718

Milligan, I. (2012). Mining the "Internet Graveyard": Rethinking the Historians' Toolkit. *Journal of the Canadian Historical Association*, *23*(2), 21. https://doi.org/10.7202/1015788ar

NDSA. (2014). *2015 National Agenda for Digital Stewardship*. National Digital Stewardship Alliance. Retrieved from http://hdl.loc.gov/loc.gdc/lcpub.2013655119.1

Pearson, D., & Webb, C. (2008). Defining File Format Obsolescence: A Risky Journey. *International Journal of Digital Curation*, *3*(1), 89–106. https://doi.org/10.2218/ijdc.v3i1.44

Peres, R. (2010). Innovation diffusion and new product growth models: A critical review and research directions. *International Journal of Research in Marketing*, *27*(2), 91–106.

Rogers, E. M. (2003). *Diffusion of Innovations, 5th Edition* (5th Edition edition). New York: Free Press.

Rosenthal, D. S. (2010). Format obsolescence: assessing the threat and the defenses. *Library Hi Tech*, *28*(2), 195–210.

Rothenberg, J. (1999). *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation. A Report to the Council on Library and Information Resources.* ERIC. Retrieved from http://eric.ed.gov/?id=ED426715

Ryan, H. (2014). Occam's Razor and File Format Endangerment Factors. In *Proc. International Conference on Digital Preservation (IPRES)*.

Ryan, H., Graf, R., & Gordea, S. (2015). Human and machine-based file format endangerment notification and recommender systems development. In *Proc. International Conference on Digital Preservation (IPRES)*.

Srinivasan, V., & Mason, C. H. (1986). Nonlinear Least Squares Estimation of New Product Diffusion Models. *Marketing Science*, *5*(2), 169–178.

Sultan, F., Farley, J. U., & Lehmann, D. R. (1990). A Meta-Analysis of Applications of Diffusion Models. *Journal of Marketing Research*, *27*(1), 70–77. https://doi.org/10.2307/3172552

The UK Web Archive. (2013). Format Profile - JISC UK Web Domain Dataset (1996-2010). https://doi.org/10.5259/ukwa.ds.2/fmt/1

Van den Bulte, C. (2002). Want to know how diffusion speed varies across countries and products? Try using a Bass model. *PDMA Visions*, *26*(4), 12–15.

Vermaaten, S., Lavoie, B., & Caplan, P. (2012). Identifying Threats to Successful Digital Preservation: the SPOT Model for Risk Assessment. *D-Lib Magazine*, *18*(9/10). https://doi.org/10.1045/september2012-vermaaten

Waters, D., & Garrett, J. (1996). *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information.* ERIC. Retrieved from http://eric.ed.gov/?id=ED395602

Wong, D. H., Yap, K. B., Turner, B., & Rexha, N. (2011). Predicting the Diffusion Pattern of Internet-Based Communication Applications Using Bass Model Parameter Estimates for Email. *Journal of Internet Business*, (9), 26–50.