A Smart Guide to Preferred Formats

Sharing knowledge and policies effectively via linked data

Sam Alloing
KB, National Library of the Netherlands
The Netherlands
Sam.Alloing@kb.nl
0000-0002-1254-1483

Remco de Boer
ArchiXL
The Netherlands
rdeboer@archixl.nl
0000-0002-9664-8500

Marjolein Steeman
Netherlands Institute for Sound and Vision
The Netherlands
msteeman@beeldengeluid.nl
0000-0002-1506-1581

Abstract – The Dutch Digital Heritage Network has developed an online tool that helps archives and heritage organisations to formulate their policy on file formats. It builds a knowledge base, via a smart combination of codification of formats on one hand and personalisation of tailor made policies on the other. This paper will explain how the online tool is set up, what technologies it uses and what possible next steps might be in the near future.

Keywords – preferred formats, community, knowledge base, linked data, registers.

Conference Topics – Exploring the New Horizons; Building the Capacity & Capability.

I. Introduction

Institutions such as archives and heritage organisations encounter many different types of file formats, some of which are more preferable than others. These organisations therefore typically formulate policies on how to treat different file formats. Formulating a file format policy leads to questions such as: *"Which formats are suitable as an archive format?"*, *"Which technical properties are characteristics of the chosen format?"*, or *"How do you create a tailored policy for your archive?"*. These are questions that every institution has to deal with, and they are pre-eminently a theme to work on together. Our approach does this by bringing all relevant characteristics of formats together in one place, by reusing as much information as possible from existing sources, and by offering a simple step-by-step plan for drawing up policy so that we don't have to keep reinventing the wheel.

The Dutch Digital Heritage Network [1] has developed an online tool called the *Guide to Preferred Formats* [2] that helps archives and heritage organisations to formulate their policy on file formats. The tool comprises a knowledge base of file formats, a classification framework that combines file format characteristics with characteristics particular to the institution, and a community portal through which institutions can register and share their policies.

This paper reports on the construction and use of the Guide to Preferred Formats. Since its public release in March 2021, the user community has been growing and several Dutch heritage institutions are using or have expressed interest in using the tool. While the tool is currently aimed at Dutch institutions, we are convinced that the philosophy behind the Guide and the way it has been linked to the international body of knowledge on file formats makes it relevant to an international setting as well.

II. Knowledge Development

To help institutions formulate a file format policy, the *Guide to Preferred Formats* combines two complementary knowledge management strategies [3]:

1. Codification, i.e. the construction of 'knowledge objects' that are stored in a knowledge base, which can be consulted to find answers to questions such as *"Which technical properties are characteristics of [a particular file format]?"*, and
2. Personalization, i.e. allow people from different institutions to engage in a dialog when they encounter similar questions such as *"What are suitable preservation strategies for [a particular file format]?"*

To this end, the Guide is composed of two main knowledge areas, which are further explained in Section III:

- The Register is the core knowledge base of the Guide, and currently contains information on approximately 1900 file formats. The primary focus of this knowledge area is the codification of knowledge about file formats so that the knowledge can be reused.
- Additionally, institutions can record and publish information about their format policy linked to formats from the Register on their own Institution page within the Guide. The primary focus of this knowledge area is the registration of 'who does what' and 'who knows about what', so that institutions can get in contact.

17th International Conference on Digital Preservation
iPRES 2021, Beijing, China.

DOI: 10.1145/nnnnnnn.nnnnnnn

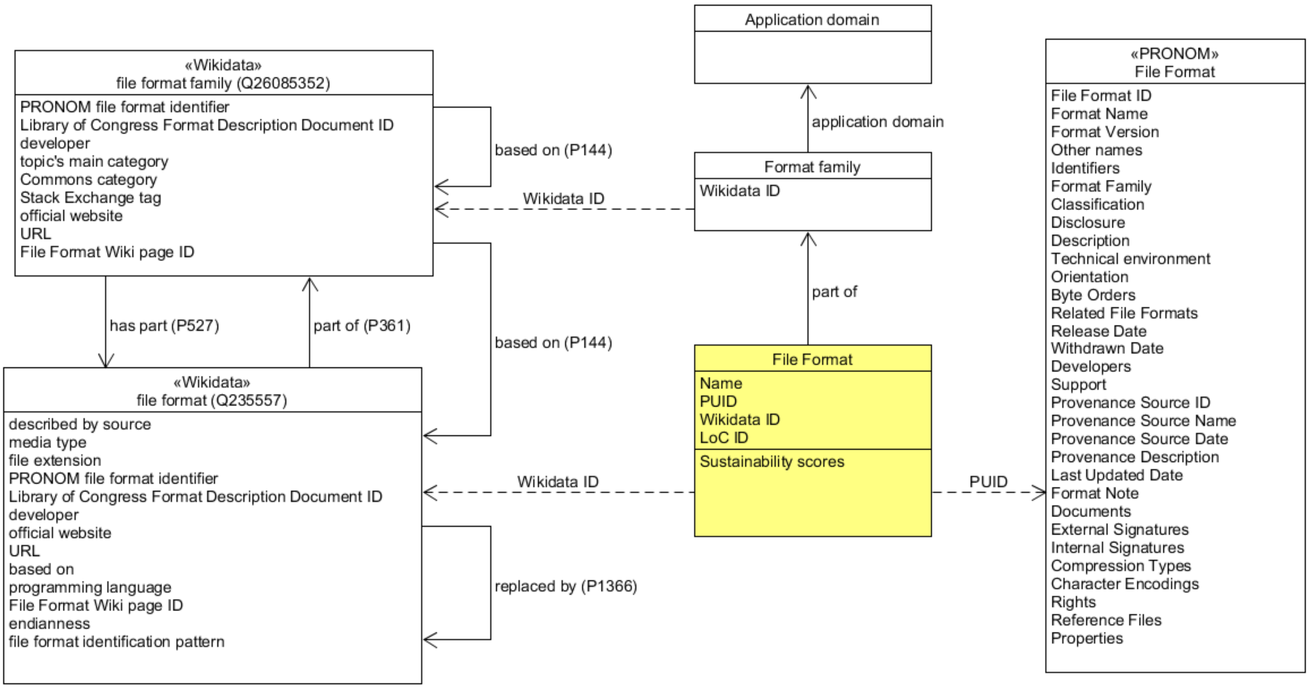


Figure 1: Structure of the Register

The two areas are connected via a framework that combines information from the Register – in particular about the sustainability of the format – with information from the Institution – in particular the importance of the format to the Institution. This framework guides institutions through a simple step-by-step plan which helps them to formulate their own policy and determine which formats should be included in their format policy.

The formats that an institution has included in its policy are displayed on its own Institution page. Institutions can also view each other's policy decisions. This can be done via the Institution page of the other institution, but also via the file format information in the Register, where you can see which institutions have already included the format in their format policy. Institutions may also include links to policy documents such as [4] that incorporate and further build upon these decisions.

The type of policy decisions may differ between institutions. When determining the policy for a file format, we often think of the qualifications "preferred format, accepted format or unsupported format". In many situations, however, that does not suffice. Often, an institution cannot prescribe which formats it will or will not receive. Institutions may also choose to first expand their knowledge about a particular format before declaring it the preferred one, in which case the *knowledge* the institution has about the format is much more relevant than the 'acceptability' of the format. After all, with little knowledge, the possibilities for sustainable preservation will also be less. It is important to let the outside world know as part of the format policy to what extent the institution is familiar with the format, and if it has for example tooling to apply identification and validation. We have therefore included knowledge levels (see also [5]) as a type of policy decision.

Where one institution has an open format policy with very few restrictions, another may be much stricter in which material it accepts. In both cases, the archive can document their format policy in our Guide. By offering institutions the opportunity to record these two aspects (or one of both), we aim for an open knowledge base about file formats in the Netherlands and hope to encourage institutions to help and find each other.

III. Building Blocks

In the previous section, we described the aim of the Guide in terms of knowledge sharing and support for policy development. In this section, we further detail the building blocks that the Guide consists of:

- The Register, a knowledge base of file formats,
- The classification framework that combines file format characteristics with characteristics particular to the institution, and
- The community portal through which institutions can register and share their policies

A. *The Register*

The Register is the core knowledge base of the Guide. It currently contains information about approximately 1900 file formats. These file formats have been grouped in file format families, which again have been grouped in application domains. The hierarchy helps the user to browse the register without having to directly deal with the more than 1900 formats. Users of the Guide can browse the hierarchy to find relevant file formats (e.g., application domain Audio, file format family AIFF contains three versions of the Audio Interchange File Format). A *drill down* interface further enhances the browsing experience, and can be used to find formats with (combinations of) particular characteristics, e.g. all file formats for a particular MIME-type.

Fig. 1 depicts the structure of the Register. Each format is represented by a separate knowledge object in the knowledge base, and is semantically annotated with relevant properties for that format. These include the name of the file format, the format family that the file format is a part of, sustainability scores (that will be further explained in Section B.1), and several identifiers that link the file format to other repositories with further information about the format. The data from PRONOM [6] serves as a starting point for the contents of the register. For each format in the knowledge base, the PRONOM Unique Identifier (PUID) is used to query the PRONOM registry for details about the format. Additionally, the Wikidata ID is used to link the format page in the knowledge base to information about the file format in Wikidata. Via a SPARQL query, information from Wikidata is incorporated in the page and directly shown to the user of the register.

B. *The Classification Framework*

The classification framework combines information about the sustainability of a file format with the importance of a file format family for the institution. This framework is presented in a step-by-step 'wizard' that guides users to formulate their own policy. The wizard offers an overview of the file format families in the Register, and provides an easy way to select the ones that are of interest. In the next step, the user defines the importance of the relevant format families from different perspectives. The importance scores from the institution combined with the sustainability scores from the Register are visualised in a plot (see Fig. 2) that gives the institution an indication of which formats should be included in the format policy: the higher the importance, and the higher the sustainability score, the more likely the file format is to be considered a 'preferred format' and/or a format that is worthwhile to build up further knowledge about.

1. *Sustainability score*

The file formats in the register contain a sustainability score, based on risk scores for several sustainability aspects. The sustainability aspects come from the U.S. National Archives and Records Administration (NARA) [7]. To calculate the risk scores, we've used the Risk matrix [8] developed by NARA. This score combines a number of criteria that influence the possibilities of preserving a format and thus calculates a weighted sustainability score. In the Guide (see Fig. 3), the criteria have been translated into the Dutch context and the scores have been adopted where possible. The scores for some components have been filled in with an expert opinion. The calculation of the sustainability score can be fully verified on the website; the approach of the sustainability score is an objective criterion that can serve as a starting point for the format policy of each institution.

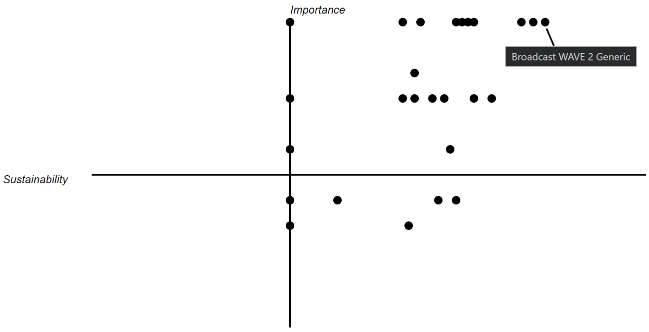
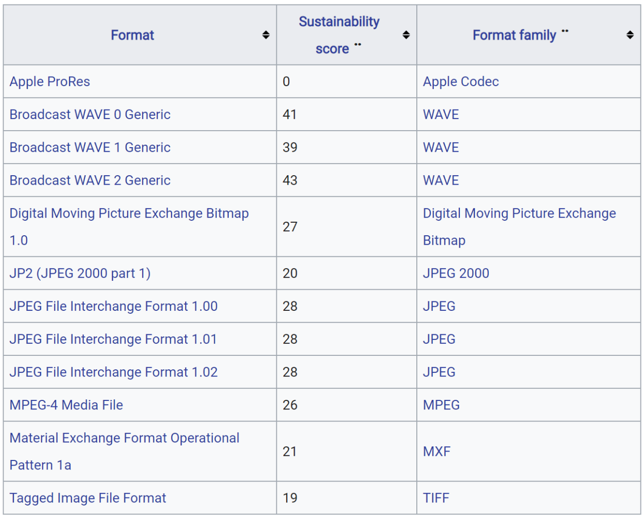


Figure 2: A plot based on the combination of file format sustainability (x-axis) and the importance of families of file formats for an institution (y-axis). Every dot in the plot represents a file format. The tooltip shows the location of the Broadcast WAVE 2 Generic file format in the plot.



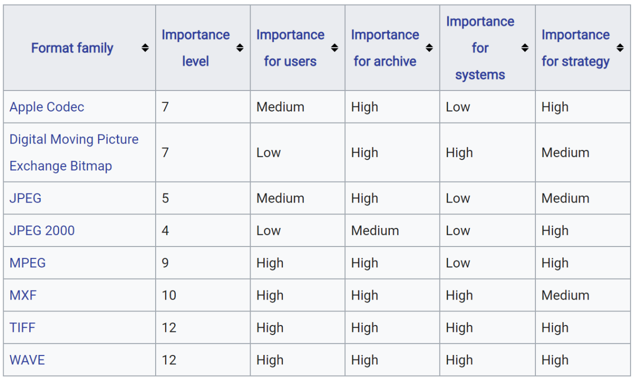
Format	Sustainability score	Format family
Apple ProRes	0	Apple Codec
Broadcast WAVE 0 Generic	41	WAVE
Broadcast WAVE 1 Generic	39	WAVE
Broadcast WAVE 2 Generic	43	WAVE
Digital Moving Picture Exchange Bitmap 1.0	27	Digital Moving Picture Exchange Bitmap
JP2 (JPEG 2000 part 1)	20	JPEG 2000
JPEG File Interchange Format 1.00	28	JPEG
JPEG File Interchange Format 1.01	28	JPEG
JPEG File Interchange Format 1.02	28	JPEG
MPEG-4 Media File	26	MPEG
Material Exchange Format Operational Pattern 1a	21	MXF
Tagged Image File Format	19	TIFF

Figure 3: Sustainability scores for various file formats. Row 4 shows the Broadcast WAVE 2 Generic format also depicted in Fig. 2.

2. *Importance levels*

While the sustainability score can be seen as an intrinsic characteristic of the file format, the importance of a file format depends on the context of the institution. An institute for Sound and Vision will consider other types of formats more important than a National Library, simply because the institution's target audience and archival material is different.

We've applied a simple model to determine the importance of a file format for an institution (see Fig. 4). We determine the importance for a file format family,



Format family	Importance level	Importance for users	Importance for archive	Importance for systems	Importance for strategy
Apple Codec	7	Medium	High	Low	High
Digital Moving Picture Exchange Bitmap	7	Low	High	High	Medium
JPEG	5	Medium	High	Low	Medium
JPEG 2000	4	Low	Medium	Low	High
MPEG	9	High	High	Low	High
MXF	10	High	High	High	Medium
TIFF	12	High	High	High	High
WAVE	12	High	High	High	High

Figure 4: Importance levels for various file format families, based on the high/medium/low assessment from the institution. Row 8 shows the WAVE family that contains the Broadcast WAVE 2 Generic format also depicted in Figs. 2 and 3.

and apply the score for the family on all formats within that family. In this way, the institution does not need to treat all file formats individually which significantly reduces the required effort. The overall importance level is determined by a combination of High / Medium / Low scores on four characteristics:

1. The importance of the file format family to the designated community;
2. The importance for the institution, with respect to responsibilities the institution has;
3. The importance from the (lack of) support for the file format family from archival systems;
4. The strategic importance of the file format family towards the future.

C. *The Community Portal*

The Community Portal is the part of the Guide that contains the information of the Institute, including contact details. On each Institution Page of the Community Portal the format policy is presented with the different aspects of the Classification Framework. The information is presented in a table and can be downloaded in CSV and used outside the Portal. The plot diagram shown in Fig. 2 shows the summary of the Framework. This information supports the knowledge exchange between similar institutions, for example Archives or Libraries, or between institutes with basic and expertise knowledge about specific file formats, for example with AV material. Section II describes this Personalisation Knowledge Management Strategy.

IV. Technology

The Guide is built as a semantic wiki using Semantic MediaWiki (SMW) [9]. The essence of a wiki is that it is a website that can collaboratively be edited [10]. In our case, members of the Expertise Group (cf. Section V) maintain the information in the register, and institutions are able to edit and add information about their file format policies and their use of the classification framework on their Institution page (cf. Section III). The essence of a **semantic** wiki is that the information in the wiki is not just stored as plain text, but is semantically annotated with properties and relations so that it becomes machine readable and queryable. The structure of the Register is itself stored in wiki pages, which makes the data model very flexible and easily extendable.

The openness of the semantic wiki platform provides all sorts of integration possibilities. SMW provides a dedicated query language that can be executed via API endpoints to query the information from the Register and the Institution pages, which could then be reused by other parties. Internally, the Guide uses the same type of queries to present information to the users of the Guide. Additionally, the data can be represented in RDF, a standard model for data interchange on the web [11]. The use of standards makes it possible to integrate the Guide with other websites as described in Section III.

V. Next Steps

The Guide was developed by the Expertise Group Preferred Formats, under the umbrella of the Dutch Digital Heritage Network. The Expertise Group has not only developed the Guide, but also provides ongoing support to the Guide. This ensures maintenance of the building blocks and the support of the community in using the Guide. The Expertise Group will be a help desk for questions about the Guide. For support on using the Guide, the Expertise Group will organise workshops and encourage community engagement, leveraging the variety of existing tools of the Dutch Digital Heritage Network.

The Expertise Group put a lot of effort in creating a Register that has a comprehensive structure of format families and versions per format. Adding and combining information from different sources does make it furthermore a rich source. From a preservation perspective it will be highly useful to assign for each format the minimum set of attributes that one should know to be in control of the significant properties of a format. These attributes probably are in part dependent on the application domain (audio, video, text etc), but may also differ per format family. It would be very useful to share and expand our knowledge on this as a next feature via the new platform.

Reusing and combining information from different sources also brought to light several limitations, for example missing data that makes it difficult to link different sources together. Where necessary, we have added this information to our knowledge base which gives us the opportunity to contribute back to the different sources and improve these sources as well.

Because the Guide uses Semantic Technologies, as described in Section IV, it is easier to expand the number of information sources even further or other sources can use the Guide as input. We will be looking at incorporating information from COPTR [12] as this is a Semantic Wiki as well. This will give us the opportunity to link Tools to the file formats. With this additional information the Guide can be used as a source for a PAR registry [13].

References

[1] *Dutch Digital Heritage Network (NDE)*. [Online]. Available: `https://netwerkdigitaalerfgoed.nl/en/`.

[2] *Guide to Preferred Formats*. [Online]. Available: `https://www.wegwijzervoorkeursformaten.nl/index.php/English`.

[3] M. T. Hansen, N. Nohria, and T. J. Tierney, "What's Your Strategy for Managing Knowledge?" *Harvard Business Review*, vol. 77, no. 2, pp. 106–116, 1999.

[4] A. de Jong, "Digital Preservation Sound and Vision: Policy, Standards and Procedures," Netherlands Institute for Sound and Vision, Tech. Rep., 2019. [Online]. Available: `https://publications.beeldengeluid.nl/pub/679/`.

[5] National Library of the Netherlands (KB), *2019-2022 Preservation Plan*. [Online]. Available: `https://www.kb.nl/sites/default/files/docs/preservation_plan_2019-2022.pdf`.

[6] The National Archives, *PRONOM - online registry of technical information about file formats*. [Online]. Available: `https://www.nationalarchives.gov.uk/PRONOM`.

[7] L. Johnston, "Creating a Holdings Format Profile and Format Risk and Digital Preservation Prioritization Matrix at the National Archives and Records Administration," in *15th iPres International Conference on Digital Preservation*, Sep. 2018.

[8] National Archives and Records Administration, *The NARA Risk and Prioritization Matrix*. [Online]. Available: `https://github.com/usnationalarchives/digital-preservation#the-nara-risk-and-prioritization-matrix`.

[9] The SMW Project, *Semantic MediaWiki*. [Online]. Available: `https://www.semantic-mediawiki.org/`.

[10] B. Leuf and W. Cunningham, *The Wiki Way: Quick Collaboration on the Web*. Addison-Wesley, 2001.

[11] R. Cyganiak, D. Wood, and M. Lanthaler, "RDF 1.1 Concepts and Abstract Syntax," W3C, W3C Recommendation, Feb. 2014.

[12] *Community Owned digital Preservation Tool Registry (COPTR)*. [Online]. Available: `https://coptr.digipres.org/index.php/Main_Page`.

[13] *Preservation action registries (par)*. [Online]. Available: `https://parcore.org/`.