PuQI - A Smart Way to Create Better Data

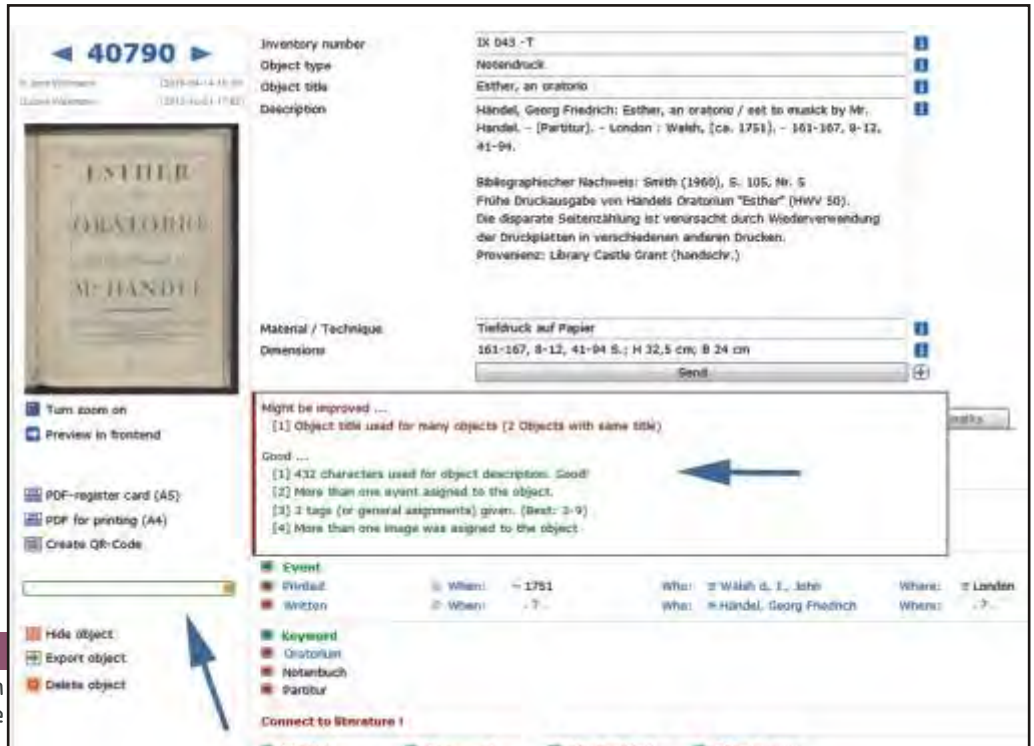
Stefan Rohde-Enslin
Germany

What does PuQI stand for?

PuQI was a city in Hubei province in China, renamed as Chibi City in 1998 by the state authorities. The PuQI this text is about has nothing to do with this city; instead, it is used as an abbreviation for "Publication Quality Index". The PuQI of this paper is a small software tool based on a mathematical formula and presenting itself as a bar and a ball. With PuQI, cultural heritage institutions get information about the publication quality of their data and they get hints on how to improve it. PuQI was developed in the context of museum-digital, but the formula and the concept behind it might easily be adapted to all kinds of software made for the administration of cultural heritage objects. Before explaining the concept and the way PuQI works, it is worth taking a closer look at the context in which it was developed and implemented. This look will reveal which elements of PuQI are specific to its current implementation, and which elements are easily adapted to other software.

The context of PuQI

PuQI is a small part of the museum-digital software. Museum-digital (www.museum-digital.de) is an initiative that started in 2009 with the aim of making it as easy as possible for museums to publish object-information on the web and to easily export them to portals of all kinds including Europeana. Right from the start, it was the aim to publish object information as effective as possible. Search engine optimization has a high priority for museum-digital. Aside from this, it has always been the aim to publish meaningful data (not only "meaningful" for search engine robots but also for humans). The objective is not to collect as many data as possible but to present data as well as possible, where "well" means "useful for researchers and the broad audience the Internet offers but also effective for search engine robots". To reach these goals there are a lot of links connecting one object to another and links that group many objects into meaningful groups. The more museums partici-



A happy PuQI turned green

PuQI translates numbers and figures into messages the museums understand and comply with.

pate with more objects, the more such connecting sites are automatically produced – but, as stated before, quality of data has a higher value at museum-digital.de than quantity of data. There is another aspect of museum-digital that is important in the context of PuQI; right from the start, all participating museums had a say in the creation and development of the software. So far, most of the wishes coming from the museum-people could be realised. Some functionalities were created because museums wanted them, but had to be moved to the background (i.e. are only visible or active for those that turn them on) – because most of the other museums considered these functionalities unnecessary or disturbing.

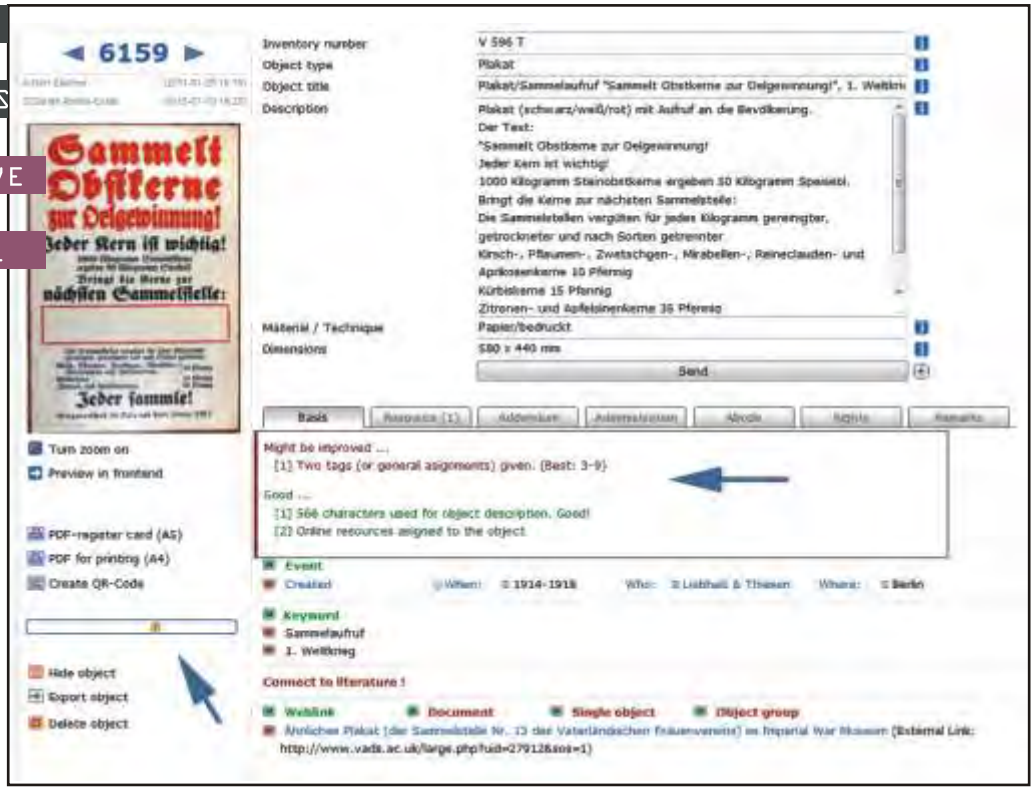
Unlike aggregators like Europeana, in museum-digital each museum has direct access to each of its objects at any time. It can enrich, correct, or delete an object, change the associated images, or append new rights information whenever necessary.

Since 2009, museum-digital has grown in numbers and in possibilities. At the moment there are more than 400 German museums participating – i.e. having at least one object online – and there is a growing number of museums who use museum-digital not only as a publication tool but also for object documentation (the respective parts of the program were introduced in 2012). Recently museum-digital became multilingual in back- and frontend and is now running successfully not only in Germany but also in Hungary. A Polish, a Brazilian, and an Indonesian version are being developed at the moment. The German versions (some are hidden since, for example, church archives are a bit hesitant with publication of objects) together administer information of about 120,000 objects of which about 58,000 are published, while the Hungarian version administers some 22,000 objects with publication of about 18,000 objects.

A fundamental principle for museum-digital is the assumption that publication and administration of museum-objects (and related information) are two very different (but closely related) things. Publication is made for public use (and search engine optimized) while administration is made for internal use. For example, for administrative purposes it might suffice to call all photographs simply "photograph" and to describe them with abbreviations and cryptic notes like "b/w" and "Child, sitting". For the use in publication, it makes much more sense to write full sentences and give names that make one photograph distinguishable from another, calling the object in this example "Photograph of a child sitting" and describing it as "Black and white photograph which shows a child that is sitting in a rocking chair at the side of a fireplace. A puppy is sitting at her knees. The photograph was taken in a studio setting...". Similar differences exist when it comes to images showing the object. For administration, a small photograph taken with a smartphone and showing the inventory number and a colour bar with the object is good enough. For publication, a bigger photograph without visible inventory number and without colour bar but with some aesthetic value is far better suited.

The range of PuQI

This is where PuQI comes into play. PuQI is not an index measuring the whole set of information assigned to (or missing from) an object in museum-digital. Only those pieces of information that are considered relevant for publication are taken into account. Because both textual and visual information are published, both are taken into account.



PuQI with hints

The limitations of PuQI

The idea of creating something that gives the museums direct feedback about publication quality first came up in 2010. There was no attempt to program a corresponding piece of software because most of the museums had the opinion that "too much control is never a good thing". After a period of abstinence from the idea, it came up again in 2013 and the museums agreed to give it a try. This time the software was programmed, implemented in 2014 into museum-digital and in the end taken very positively by the museums. It was made sure that every museum can only see the index for each of its own objects. All museums know about the limitations of PuQI and importantly, they know that all the index is giving them are indications. There is nothing forcing them to follow the proposals of PuQI. Not forcing anything but giving advice has proven to be a good concept.

The software behind PuQI is very small (a script of some 390 lines) and is limited to counting and interpreting numbers. PuQI cannot measure intellectual quality or correctness. For example, it is not able to detect if a wrong image is attached to the object information, but rather measures the number of images attached and the size of each. PuQI is in use but at the same time it is continuously being improved on the way towards an answer to the question, "How far can quality of object-publication-information be measured and improved by counting numbers alone?"

The PuQI way

As said, all PuQI knows are numbers, or to be more specific, the number of words used at a certain place, the number of letters used at another, the number of images, and the sizes of images. For example, PuQI in its current implementation measures whether "object measurement" information is given or not simply by detecting if the number of letters in the field "object measurement" is zero or higher. Another field PuQI takes into account is "object name". Here PuQI measures whether the name consists of more than only one word or not and it checks how many objects with the same proposed name are already known. Each field and state then has a marker of importance and rules assigned to it; for example, missing measurement information gives minus 3 points or an object name consisting of only one word gives minus 3 points. Repetition of "object name" will result in minus 4 points. So a set of object information without measurements (-3) and at the same time without information regarding material /

technique (-3) and with an object name consisting of only one word (-3) which at the same time is the name of other objects (-5) would give a quality score of (-3-3-3-5=-14 points). Considered most important for publication is the object description (search engines will like it). Here again, PuQI measures the number of letters according to the following rule: 1-49 letters is far too short (-25); 25-49 letters is still too short, but ok (-10); 50 to 249 letters is considered as short (-5); 250-899 letters is considered good (+3) while 900-1746 letters is very good (+6); 1750-4999 letters is seen as probably too long but ok (0), and more than 5000 letters is not suitable for a web page – it might be better to upload a document and attach it to the object (-6).

The museum-digital software is built with LIDO (in mind, which is why it knows "events" (what happened to the object when, where, and by whom). For every such "event" a museum attaches to the object, the index will rise. Museum-digital also works with tags attached to the objects (used in the sense of labels for topics or themes or contexts the object belongs to). For each tag, the object gets points until the number of nine is reached. Too many tags will probably confuse the website visitor, so for each tag above 9 (the 10th, 11th …) 3 points are subtracted.

For each image, audio, or video attached to the textual object information 3 points are given – except when the resource is too small, in which case points are subtracted. Missing rights information will result in -15 points for each resource – if neither rights holder nor rights status is given. So an object with ten images attached of whom two are too small and three are without any rights information will get -19 points – however nice and pretty the images are (10x3)=30 basis points for images, 30-(2x-2)=26 subtraction for size, 26-(3x-15)=-19 subtraction for rights information – or: (10 x 3) + (2 x -2) + (3 x -15)=-19). That is basically the way PuQI works.

PuQI in a nutshell (the principles)

1) Select the fields that should be considered;
2) Rank these fields according to relevance for good and effective publication;
3) Define rules for each of the fields (good, not so good, bad);
4) Quantify the entries in the fields and assign them a status (good, …) expressed by a number (e.g. -5) where the number depends on the ranking of the field;
5) Count all the values you get and compare the total to the "reasonable maximum" (the number one gets putting "good" in all fields).

The basic question here is: What is a good entry? This has to be defined field by field. There cannot be a general rule in force for all of the fields, and it depends on the purpose. Optimizing for Google alone is one thing; for a special portal like Europeana is another. Here it would be best to define "good" more generally, i.e. good for search engine robots, and for users and for portals of all kinds. There has been only limited discussion about the question of "what is good under what conditions in which setting". There should be an intensive discourse about it.

PuQI - an example implementation

To give an example (and as an invitation to discuss it) the following table shows how in museum-digital the "good entry" per field is defined (with good publication for all purposes mentioned in mind). The values and rules were created through experimentation.

Field	Ranking	Rule	Remarks
Object type (This is a field required in LIDO)	-	If more than 2 words: Create a warning message only	The object information cannot be stored if there is no information about the kind of object. But, usually "Object type" should be a one-word-term and not a whole object title
Object name (The title for an object)	Middle (-5 to 0)	If only one word: -3 If more than ten words in the title: -5 If the object name is already used for other objects: -5 Else: 0	Two checks are done: a)Number of words (one is too short and ten is too long) b) how often the same object name is applied
Object description	High (-25 to +6)	If less than 25 characters: -25 25 to 49 characters:-10 50 to 249 characters: -5 250 to 899 characters: +3 900 to 1749 characters: +6 1750 to 4999 characters: +0 5000 and more characters: -6	Two checks are done: a)Number of characters and b)if same description is used for other objects
Material / Technique	Low (-3 to 0)	If no word or character: -3 Else: 0	Check only if there is an entry
Measurements	Low (-3 to 0)	If no word or character: -3 Else: 0	Check only if there is an entry
Assigned to a collection	Middle (-10 to 0)	If not assigned to a collection: -10 Else 0	Check only if there is an assignment
Events	High (-15 to +5)	If no event is assigned: -15 For each assigned event: +5 If the place of the event is assigned additionally as a subject: -15 for each assignment), Same for actors that take part in the event and are assigned again as a subject : -15 for each assignment). Same for time-entries that take part in the event and are assigned again as a subject : -15 for each assignment)	Number of events is measured and it is Check for unnecessary information. If a place/actor/time is part of one of the events related to the object there usually is no need to assign an additional general relation in the subject fields

Tags	High (-15 to +27)	If no tag is assigned: -15 If one tag is assigned: -10 If two tags are assigned: -5 For tag number if there are three to nine tags: +3 for each tag For each additional tag -3	Check the number of assigned tags. More than ten tags will confuse visitors and search engine robots (we got a mail from Google!)
Literature	Low (0 to infinite)	If literature is assigned: +3 for each assigned literature Else: 0	Check if literature is assigned and respect number of assignments
Weblinks	Low (0 to infinite)	If links are assigned: +3 for each assigned link Else: 0	Check if weblinks are assigned and take number of weblinks into account
Documents	Low (0 to infinite)	If documents are uploaded for the object: +3 for each document Else: 0	Check if documents are uploaded and take into account the number of documents
Object-Object-Relation	Low (0 to infinite)	If the object is related to other objects: +3 for each listed relation	Check if relations to other objects inside the database are stored and take number of relations into account
Object-Objectgroup-Relation	Low (0 to infinite)	If the object is set into relation to objectgroups: +3 for each listed relation	Check if relations to object-groups are stored and take into account the number of relations
Images / Resources	High (-infinite to infinite)	For each image basically: +3 For each image (additional counting): - Long side less than 600px: -5 - Long side 500 to 800px: -2 - Long side more than 800px: 0 No image owner but rights status given: -10 No rights-status but image owner given: -10 No owner AND no rights-status: -15	The number of images or resources (i.e. movies, audio records, …) is measured and the size of the images is checked (in museum-digital it is impossible to publish an object without any image) as is also rights information

That's it. The small PuQI script does the necessary checking and counting and produces for every state of a field (if necessary) a message. After formulating the messages, the script sorts them according to types. There are three types of messages:
- "General remarks", messages of this type are presented in blue letters
- "Might eventually be improved", messages of this type are presented in red letters
- "Well done", messages of this type are presented in green letters

The PuQI messages

A "General remark", for example, is formulated if the number of words in the field "object type" is higher than two. In this case the message might look like "3 words in field 'object type' – is this really what is meant? (Best: One word)". All remarks avoid terms like "error" or "mistake" or "false"; rather, they propose thinking twice and they clearly state what would be good or best.

A "Might be improved"-message looks like "203 characters used for object description. That is quite short!" – a clear indication.

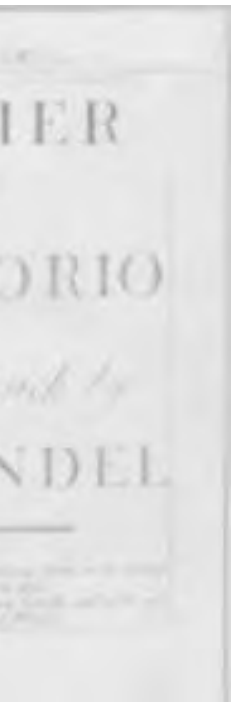
Finally, a "Well done"-message looks like "3 tags (or general assignments) given (Best: 3-9)".
All the messages are formulated very carefully and in the spirit of not forcing anything, but rather to inspire and show the best way: Benevolent guidance, which one does not have to follow but - as reality demonstrates – is taken seriously.

The PuQI appearance

Equally important for the acceptance of PuQI by the museums is the way it is designed: Graphically, PuQI is a simple bar with a ball. Bar and ball change colour according to the index. If the index is very low, both are coloured red; if it is very high, both are coloured green. If the index is in-between, bar and ball turn blue. Apart from this, the ball goes from left (low index) to right (high index).

PuQI numbers were experimentally adjusted to reality and tuned for motivation. Intuitively, one might think that a ball on the left edge of the bar (in red of course) equals an index of zero, and a ball on the right edge (in green) symbolizes the highest index. This is not the case: The left / red section is kept quite small; the medium / blue section is a bit larger; but the right / green section is the largest. (Index below -30 is red, Index between -30 and 9 is blue and everything above is green). The outcome is that it is quite simple to turn line and ball into green – which is understood as a confirmation by the colleagues using the system. They see that they are on the right way. At the same time, it is made difficult to move the ball to the very right – where the ball only appears if some extras (e.g. 8 tags for the object and 5 images – including rights information - assigned) are there. If all fields are filled in / used without extras (which means that all that should be there is present), then the (green) ball will reach only 4/5 of the line. The whole line goes 25 % above "good" (or "reasonable maximum"). Observation shows that, because it is quite easy to reach the green

The PuQI



colour, the colleagues try to avoid having any other colour for the PuQI bar and ball. It also shows that the more they use the system, the more they try to move the ball to the right edge of the bar. Many colleagues even stopped putting new objects into museum-digital for a while, and instead were reworking all their existing entries according to the suggestions of PuQI.

PuQI -Graphics and numbers

The whole PuQI only works with numbers and counting, but those entering the data into museum-digital don't see their index as a number - only in the form of a coloured bar and ball. It would have been easy to give them a number, e.g. saying "The PuQI of this object currently equals 34" – but this would have misled some colleagues who, through all means possible, would have tried to reach 100 or even more, spending too much time on improving information on one object instead of entering new objects. Other colleagues might stop entering any new objects at all, because with the information they have, an index of 100 could never be reached.

PuQI the unavoidable

One last thing about PuQI: It works with a so-called mouse-over effect. If one touches the bar or ball with the mouse, a window pops up showing the messages. The placement of PuQI inside museum-digital is such that the "Publish object" button is right below the PuQI line, so that it is very hard to avoid PuQIs messages: the mouse will cross the line on its way to the button …

That's all there is to say about PuQI. Some parts of the way it is implemented are specific for museum-digital, but in most parts PuQI (or something similar, build on the same principles) might easily be adopted and integrated into all museum-software-programs.

It is easy to follow the principles of PuQI (definition of relevant fields, ranking them, setting rules for all relevant fields, and calculating an index according to these rules in combination with the ranks of the fields) to create other indexes, e.g. a Museum-object-documentation-Quality-Index (ModQI).

PuQI - Quality by quantifying?

When programming the PuQI the question was, "Can quality be improved by measuring only quantitative values?" Observation shows that – at least in the framework of museum-digital, where the PuQI is smart, benevolent, unobtrusive (and unavoidable) – this can be very much the case if the index is presented thoughtfully. PuQI translates numbers and figures into messages the museums understand and comply with.