Quality control tools user requirements

Nathalie Denos

Contribution to deliverables UR, WP 2 / QCT task

I. Introduction

The members of the HEP community naturally carry on article evaluation activities, in their daily work. The idea of the quality control tools is to benefit from this naturally arising activity to feed an evaluation database that contains not only the documents (records) but also various quality-related features that cannot be obtained in an automatic way. These additional document features will allow for new searching facilities, by including additional search criteria such as quality-related criteria. In the same time, they will imply and allow for more control on the quality of documents.

The actors of evaluation

Evaluation of articles by users occur in two different contexts: evaluations made by referees mandated by a journal or a conference editorial board, to decide whether or not to publish an article; evaluations made by regular users of databases who retrieve articles for their research work, and read some of them.

In the first case (*referee evaluation*), the referee must give a detailed evaluation, and must take some time to do it. Moreover, the evaluation he/she makes must be communicated to other people (the editorial board, the author himself), and must conform to the journal or conference standards for publication. Hence the form of the evaluation is explicit, argumented and objective.

In the second case (*simple reader evaluation*), the reader will evaluate the articles retrieved for his/her own use. As a consequence, he/she will not necessarily make the evaluation explicit, argumented or objective. He/she may not be willing to share this evaluation with other people, as it would require a bigger effort, and more time.

We have to distinguish these two cases.

Given that we can expect to have a detailed evaluation of articles only if they have been part of a referee process, many documents may not benefit from this evaluation. Then, we must find *incentive means* to bring simple readers to contribute to the evaluation database. The point is that simple readers are neither necessarily good nor cautious evaluators. As a consequence, *automatic processes* should be set up to *contribute* to the evaluation database, and to *control the evaluation* information provided by simple readers.

Automatic evaluation

Fully automatic evaluation of articles is not possible. Nonetheless, it is interesting to study in detail the various quality features in order to estimate their amenability to calculation. As user groups are available, TIPS is a good context to evaluate the performance of the automatic evaluation tools.

Identification and privacy

Another important topic is that some evaluations may be available to any user, whereas some are of a private nature. As a result, users who contribute to the evaluation database must be identified, and access to evaluation information must be restricted according to the identity and role of the user.

Outline

We first review *document quality features* and then analyze and discuss their amenability to calculation, especially in the context of TIPS information resources.

II. Document quality features

Many quality features have already been defined in the literature, concerning the quality of data, the quality of Web pages, and the quality of scientific literature. Some of them can be automatically calculated, with more or less reliable processes; some of them cannot.

For this reason, we first give the most exhaustive list of quality criteria that are important to the searching of scientific articles, and discuss the possibility to automate the calculation of the corresponding features, as well as the benefit that could be drawn from having readers assess them. Then we analyze them and discuss their feasibility for the PORTAL. Finally we describe in more details the features that we choose to implement in the first version of the PORTAL.

A. Overview

Before starting, we first go into definitions and examples to clarify the meaning of "quality criteria", of "quality features", and to give a quick overview of the arising issues.

Quality criteria are criteria that allow selecting a subset of documents out of a set of topically relevant documents. They allow users to specify quality standards when they search documents.

Quality features are the properties of documents that are associated to quality criteria. For a given document, a quality feature must be assessed to allow for later searching along the corresponding quality criterion.

Variety
There is a wide variety of quality features.

For instance, " correctness" is a quality feature that reflects a strict meaning of the word "quality", as it vehicles a clear-cut judgment: it is always bad for a document not to be error-free. On the opposite, "recency" is another feature that reflects a less clear-cut meaning for the word "quality", as a document may be a good one, although it is not a recent one.

So we must keep in mind that quality features as we think of them here, are not always connected to the strict notions of good and bad. They refer to features that are likely to contribute to the description of non-topical document properties. Sometimes, their evaluation will strongly depend on the user's situation.

Quality feature evaluation
When a quality feature is evaluated for a given document, a value is associated with it. This value can be viewed as the answer to a particular question concerning the document. For instance, to set the value of the quality feature "Recency" for a document, one has to answer the question "how recent is this document?".

To answer the question "how recent is this document?", it requires the value of "date of publication", in the case of a published document, and other features in the case of non published documents. This information is important as the availability of the "date of publication" will determine the possibility to calculate recency.

Document features

"Date of publication" is a feature used to assess recency. This type of feature is called a document feature, because it concerns the document and nothing else. Document features are generally available in standard bibliographic records.

Top-level quality features and hierarchical structure of feature evaluation

An interesting example of a top-level quality feature is "authority", meaning "the power to influence or command thought, opinion, or behaviour". To evaluate the authority of a document, many other quality features of the document can be used, like "reputation of the author", "reputation of the editor", etc. It can in turn be evaluated on the basis of the basic feature "author", together with the help of an expert's opinion, or a social tool able to assess the reputation of this author.

With this example, we see that quality features can be organized in a hierarchic way, going from top-level quality features down to more basic features, among which some are directly available in standard bibliographic records and others are not. The structure is given by the "requires" links between features.

Community features

"Reputation of the editor" cannot be called a document feature, as it does not concern the document itself, but the editor of the document, and his/her position in the whole research community. Reputation does not only depend on a particular user’s opinion on the editor (he/she may not have any opinion), but also and mainly on the general opinion of the community.

With this example, we see that some quality features require elements of the opinion of the community. We call this type of feature community features.

Collection features

To estimate the authority of a document, a traditional method is to use the "citation index" of the document, i.e. which documents cite the document of interest. The citation index is thus a feature that depends on the other existing documents in the field. We call this type of feature collection features, as they require the whole collection to be evaluated.

Contextual features

Under the general availability quality feature, several aspects are covered, such as the availability of the full text, or the download time. To evaluate the download time, one needs not only the size of the file to download, but also the "network load" at the moment of download. This type of feature is called "contextual feature".

User's information searching situation

Lastly, the assessment of some quality features is deeply rooted in the current information searching situation of the user. For instance, the "timeliness" of a document is definitely a quality feature of a document, but it depends on the user's current state in his/her information searching process. In other words, this feature is not well suited for the making of an objective reusable long-term evaluation.

This is the case for a number of quality features (grouped under the general quality feature called "situated usefulness"). We will only mention these features for the sake of exhaustivity, without going into details.

Multiple evaluation processes
To go on with the "intended audience" example, this feature is not a simple one, like "date of publication", for instance, as it is scarcely directly available in standard bibliographic records. To determine the value of this feature for a given document, several ways can be imagined: computer-based estimation through a specific analysis of the full text of the document, or through a search for explicit mentions of the intended audience that may appear in the title, abstract or forword, or, lastly, through the direct evaluation of a user who reads the complete document (author or acknowledged reader).

We see here that sometimes, several ways are conceivable to evaluate a quality feature, ranging from computer-based estimations to direct human evaluation. Each way is likely to require specific features.

Summary
Hence we see that to evaluate *top-level quality features*, *document features* are not always sufficient; sometimes we will need other *non top-level quality features*, *community features*, *collection features*, *contextual features*. We must add to this list the *user's current state* in his/her individual information searching process, although it will be mentioned but not developed here.

B. Description of features

In the following, we describe the top-level quality features that we have identified as useful to the searching of scientific papers are first described, with the hierarchy of features required for their evaluation. Then, we group together features into the categories identified above (document features, community features, collection features, user features, contextual features, and non top-level quality features).

1. Top-level features and hierarchy of required features

We have identified the following top-level quality features:

- Scientific quality
- Readability
- Intended audience
- Quality of identification
- Recency
- Availability
- Authority
- Popularity
- Situated usefulness

We give here definitions of the features and an indication of the required features. As we do not give here details concerning the possibilities to estimate the features when they are not directly available, we shall not mention the features that may be used for this type of estimation.

a) Scientific quality

The features that pertain to the scientific quality are features generally used in peer review.

The scientific facts presented in a document can be evaluated along the following features: "correctness" (number of errors), "completeness" (number of omissions), "accuracy", "quality of methodology", "quality of demonstration", "quality of references list", "originality", "currency".

Currency differs from recency: currency refers to the general use, acceptance or prevalence at the moment, whereas recency simply refers to the date. To evaluate currency, field knowledge is required.

Scientific quality
Correctness
Completeness
Accuracy
Quality of methodology
Quality of demonstration
Quality of references list
Originality
Currency

b) Readability

The "readability" of a document can be evaluated along the following features: "quality of writing style", "clarity of expression of ideas", "quality of logical structure", "adequation of illustrations", "absence of repetitions".

Some of these features can be estimated automatically, with heuristics, although the results may be very poor or very far from a reader's conceptions. Here again, these features, if automatically computed, must be regarded and exploited cautiously.

Readability
Quality of writing style
Clarity of expression of ideas
Quality of logical structure
Adequation of illustrations
Absence of repetitions

c) Intended audience

A scientific document's intended audience can be examined along its "technical level", its "educational level", and its "breadth of field covered", although these features are scarcely available in databases.

Moreover, a document can often be associated with a "type of research work" (experimental, theoretical), and to a "type of document" (review article, technical report, research report, conference article, course notes, PhD thesis, working notes, raw experimental data, etc.), although these features are not always available in databases. These features can contribute for a user to evaluate the adequation of the intended audience of a document, at searching time.

Intended audience
Technical level

Educational level
Breadth of field covered
Type of research work
Type of document

d) Quality of identification

The question for a user to be able to identify with certainty a document is important in bibliographic searching.

When a user wants to cite a document in an article, he/she needs the to have a complete enough bibliographic reference. The "citability" feature requires then the "bibliographic reference" feature, as well as the "publication status" feature to check whether it is complete or not.

In addition, as it occurs in the HEP field that preprints are frequently uploaded to databases, it is important for users to be able to know whether a document is a preprint of a published article, or if it is a different one. We call this feature "duplicates trackability". It implies "bibliographic reference" feature, but also the "history" and "publication status".

Another feature of an electronic document, when it is part of an electronic database, is the document identification number, or "document id.". When databases comply with the Santa Fe convention, their document id identifies a record in a unique way. It does not mean that this record is not a pre-print of another record, but during a search it can be helpful to be able to see that two records retrieved at different times, are identical.

Quality of identification

Citability	*Bibliographic reference*
Duplicates trackability	*History*
	Bibliographic reference
	Publication status
	Document id

e) Recency

The "recency" feature gives an indication of when the document has been written. The best feature is thus the "writing date", but it can also be given by the "publication date" of the document, or the "submission date" to a journal or confeerence if the document is destinated to publication. When a document is not published (preprint for instance), the "accession date" (Santa Fe convention terminology for "date of upload"), can be a useful feature too, as well as other features such allowing to infer the date when the document was written.

Recency
Writing date
Publication date
Accession date

f) Availability

When a user is about to evaluate the usefulness of a document that he/she founds online, it is important to know what of this reference is available: the full text or the abstract only, or nothing outside the reference itself. This aspect of availability is called "absolute availability".

In the context of electronic database searching, another aspect of availability is the "download time", and can be estimated on the basis of the "size of file" to download and on the "network load" features. The download time will be of course conditioned by the bandwidth available to the particular user.

Also, when a user is interested in a document, it may be important for him/her to know whether he/she will be able to retrieve it again later. The "durability" feature is thus part of the "availability", and can be based on the "anticipated future" of the document. More generally, one can base the durability feature on the electronic publisher's policy concerning the management of documents in time.

A user has tools to view and print documents on his/her computer, which may constraint his/her ability to use and read the document retrieved. This aspect reflects the exploitation ease. It requires the "formats of full text" feature to determine the "viewability" and "printability" features for a given user, who knows which viewing and printing tools are available to him/her.

Availability

Absolute availability	*Full text*
	Abstract
Download time	*Size of file*
	Network load
Durability	*Anticipated future*
Viewability	*Formats of full text*
Printability	*Formats of full text*

g) Authority

The "authority" feature refers to the confidence that can be attributed to the paper on the basis of its publication context and authors. It implies features such as the "reputation of the author", as well as the "reputation of the organization" the author belongs, the "reputation of the publisher" of the document, and the "reputation of the editorial board members" that judged the document worth published, and the "reputation of journal or conference". These features can refer to generally agreed reputations, rather than to a particular user's opinion on these reputations.

In addition, the feature "publication status" can contribute to the evaluation, as many of these reputation features do not apply outside the scope of a published document.

The "reputation of..." features have a taste of subjectivity, but there are trditional ways to give estimates of reputation with metrics accounting for "citation index", "self-citation" and already existing data on the quality of the implied documents. Of course the scores obtained with metrics can never be taken as reflecting even correctly reputation (because of non-exhaustive data available, biases, etc.). For this reason, they cannot be used as an objective basis to properly evaluate a person or an organization, but they can be useful to select a subset of too big a set of retrieved documents.

Authority

Reputation of author	*Author*
	Citation index
	Self-citations
Reputation of organization	*Organization of author*
Reputation of publisher	*Publisher*
Reputation of editor	*Editor*
Reputation of journal or conference	*Journal or conference*
Publication status	

h) Popularity

The "popularity" feature refers to the degree to which the document is actually used by people. It is connected to the "citation-index" features, as the academic citing activity gives indications of the actual academic use of documents by members of the community. As an important complement of the academic aspect of popularity, we must consider the "downloads" and the "recommandations" regarding the document, as these features may allow to capture the use of non-academically acknowledged documents that are nonetheless very much used by researchers.

Popularity
Citation index
Downloads
Recommandations

i) Situated usefulness

Eventually, the usefulness of a document strongly depends on a user's current situation and information need. More precisely, it involves his/her personal previous knowledge, and encompasses topical relevance features. We will not detail this feature as it will not be exploited later.

2. Summary by type of feature

Here is an organized list of the features involved in quality features evaluation, by type of feature.

a) Document features

Correctness
Completeness
Accuracy
Quality of methodology
Quality of demonstration
Quality of references list
Originality

Currency

Quality of writing style
Clarity of expression of ideas
Quality of logical structure
Adequation of illustrations
Absence of repetitions

Technical level
Educational level
Breadth of field covered
Type of research work
Type of document

Citability
Duplicates trackability

Bibliographic reference
- Title
- Author
- Organization of author
- Publisher
- Editor
- Journal or conference

History
Publication status
Document id.

Writing date
Publication date
Accession date

Absolute availability
Exploitation ease
- Viewability
- Printability

Ease of download
Durability
- Anticipated future

b) Community features

Reputation of author
Reputation of organization
Reputation of publisher
Reputation of editor
Reputation of journal or conference

Downloads
Recommandations

c) Collection features

Citation index

d) Contextual features

Network load

3. Overview of the hierarchy

Figure 1 gives the hierarchy of features, except those relating to situated usefulness. At this point of the presentation, we omit potential required features that may be used to automatically estimates features that are not directly available.

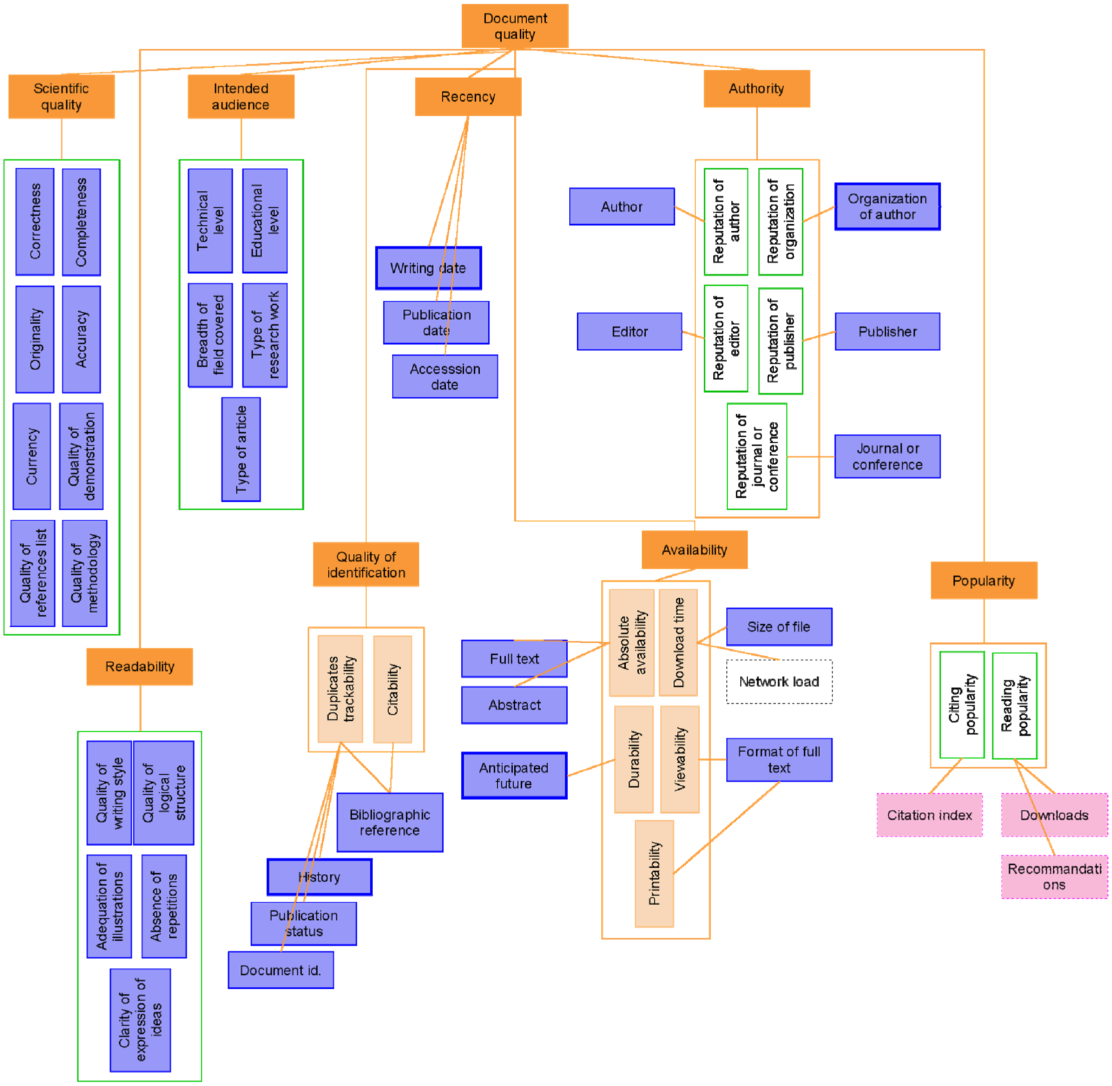

Figure 1 Hierarchy of quality features

III.Analysis and discussion for choice

We first give a summary table of the availability of document features in arXiv and JHEP bases, as they are described in the information resources document. Then we discuss in detail the availability of features in the TIPS information resources, and the amenability to calculation of non available features on the basis of the available ones.

We start with top-level quality features. For each of them, we discuss the availability of the other features required for the top-level one. When they are not available, we give indications of alternate evaluation processes, among the following possibilities:

- direct calculation
- computer-based estimation
- human evaluation

For each evaluation process, we give an indication of the *reliability* of the results obtained, and the *context* (type of user activity) in which the evaluation process can take place.

For computer-based estimation processes, the reliability of the results will depend on the a priori precision of the method and of the availability of the data on which the method can be based. This indication allows to identify those that will require an important validation phase, and to set up priorities on processes to implement in a first approach.

For human evaluation processes, the reliability will depend on the context of the evaluation and on the type of user involved in it.

A. Features availability in arXiv and JHEP

The following table summarizes the availability and amenability to calculation of features for arXiv and JHEP.

Some of the ground level features are directly available (1 means available, 0 means not available, "u" means "unknown"). Sometimes, features are available, but not for all records. In this case they are said to be available (1), but a percentage of availability is given. When this percentage is unknown, the value is "u". When they are not, the question of their amenability to calculation is asked, and scored from 0 (directly available) to 4 (impossible to calculate), with intermediate values meaning: 1 (easily calculated); 2 (difficult to calculate but likely); 3 (possibly calculated, unknown method). Value "u" means "unknown". Value "-" means not applicable (because available at 100% !).

		arXiv			**JHEP**		
		Available	Percentage of availability	Amenable to calculation	Available	Percentage of availability	Amenable to calculation
	Title	y	100	-	y	100	-
	Author	y	100	-	y	100	-
	Organization of author	y	u	u	n	0	u
	Abstract	y	100	-	y	100	-

		arXiv			**JHEP**		
		Available	Percentage of availability	Amenable to calculation	Available	Percentage of availability	Amenable to calculation
	Full text	y	100	-	y	100	-
	Publisher	y	u	u	y	100	-
	Editor	y	u	u	y	100	-
	Journal or conference	y	u	u	y	100	-
	Publication status	y	u	u	y	100	-
	Publication date	y	u	u	y	100	-
	Full bibliographic reference	y	u	u	y	100	-
	History	y	u	u	y	u	u
	Type of research work	y	u	u	n	0	u
	Type of article	n	0	u	n	0	u
	Size of file	y	100		n	100	y
	Formats of full text	y	u	u	y	u	u
	Accession date	y	100	-	y	100	-
	Downloads	u	u		u	u	

B. Discussion of features for arXiv and JHEP

For the arXiv and JHEP databases, we discuss the possibility to integrate the quality features into the PORTAL.

1. Scientific quality, readability, intended audience

It is not possible to do a computer-based estimation of these features in general, they all require human evaluation.

For JHEP papers, which are reviewed, these features could be filled up by reviewers themselves.

For arXiv, there is little chance to obtain such information, outside asking simple readers to give their own evaluations, especially for those features that do not vehicle a quality judgment (to avoid personal interest biases), i.e. for intended audience.

Intended audience
Technical level
Educational level
Breadth of field covered
Type of research work
Type of article

Readability
Writing style
Clarity of expression of ideas
Quality of logical structure
Adequation of illustrations
Absence of repetitions

Scientific quality
Correctness
Completeness
Accuracy
Quality of methodology
Quality of demonstration
Quality of references list
Originality
Currency

2. Quality of identification

Citability..*Bibliographic reference*

The citability feature depends on the completeness of the bibliographic reference.

A partial bibliographic reference can always be made up out of the Title and Author fields. The issue is to obtain a *full* bibliographic reference. It is important as it is useful to users when they want to cite articles in a paper they are writing.

For the JHEP base, the complete reference is always available, even if the paper has not been accepted for publication, as JHEP editors know this (Publication status is set to "not published", or "submitted").

For the arXiv documents, the completeness of the bibliographic reference cannot be guaranteed. Additional bibliographic elements can sometimes be found in the Comments field, and sometimes in the Notes field. For instance, one can find information such as "Talk presented at the International Symposium....", or "Contribution to proceedings of III Int Conf....", or "Lectures delivered ar the CXXX Course of the International School of Physics...", or "To appear in Int. J. Math...". The important thing here is that there is no agreed format for these pieces of bibliographic information, as they are freely entered by authors. As a consequence, it is not a simple matter to exploit it: heuristics could be imagined to identify the portions of text that give bibliographic information, but then, the issue of identifying explicitly a journal's name, for instance (in order to be able to check identity with another record) is very hard to do.

A possible solution for the arXiv this would be to ask more from authors at upload time, with the following rule: mandatory Bibref field to fill up (into a standardized format), with the implication that when a paper has no full Bibref field, it means that the paper has not been published, nor submitted (Publication status field will indicate this), which reduces considerably its authority! This can be an incentive way to have authors fill up the fields.

Duplicates trackability.....................................*History*
....*Publication status*
....*Bibliographic reference*
....*Document id.*

Duplicates trackability must reflect the ability for a given document, to be identified as being identical or very similar to another one, as, for instance, it is the published version of a preprint, identical or slightly revised. As a single paper can be submitted to several archives, this point is important, for users not to retrieve and downlod the same document several times, and to be able to identify which version is best.

Estimation of it will vary, depending on the publication status of the document. If it is a preprint, duplicates trackability is good if the author has provided the indication of the journal or conference where it has been submitted, and the date of the submission... If it is a published document, duplicates trackability is good if the author has provided the document-id of the corresponding preprint(s) (with Santa Fe convention for Open Archives), or the explicit mention that it has not been submitted anywhere as a preprint. This is what we call the *History* feature.

For JHEP, the History feature exists, in the XXXref field for the correspondance with arXiv documents. To improve the estimation of duplicates trackability, it would be good to allow for the mention of any open archive document ids, in case of multiple submission of preprints. We do not know if this field is always filled up, and if when it is not, it means the not preprint has been submitted. An additional "none" value may help to distinguish this case.

For the arXiv preprints archive, the Journal-ref field and the Report-no field give indications of where the preprint has been published. In addition, the Comments field sometimes includes an indication of the type "submitted to...", in free text form mixed with other information. The optional Notes field also bears this kind of information sometimes ("to appear in..."). A tool can be designed to extract this information from the existing free text fields, but a real improvement would be that a specific field is added, to be filled up by authors at upload time.

3. Recency

Recency is ideally based on the writing date, but it is generally not available. It can be estimated on the basis of the publication date, when the paper has been published, and on the accession date if it has not been published.

Writing date
Publication date
Accession date

In JHEP, the following dates are available: reception date, acceptation date, publication date. Accession date is not relevant for JHEP papers, as the reception date does not correspond to the accession date (date to which the document is accessible to other people). On the opposite, publication date is always available. It must be noticed that recency could be estimated using the reception date too, as it gives a more precise idea of when the paper's writing has been finished.

In arXiv, the accession date is available in field Date. It must be noticed that it is followed by the size of the submitted file, so the date will have to be extracted from this field. Publication date is relevant every time the document has been published. It appears in field Journal ref. As Publication date is not always relevant in a strict meaning, we can think of some other dates can be useful, such as, for lecture notes, the date when the lecture was given, or for a thesis, the date when it was presented, etc. These dates are sometimes available wherever bibliographic information appears (fields Comments and Notes). It is worth using these dates to fill up the Publication date feature, as it may help estimating Recency.

4. Availability

Availability encompasses several categories of quality features: the absolute availability of the data (full text and abstract); the download time; the durability; the possibility, once downloaded, to view or print it for a user.

Absolute availability

Full text availability	*Full text*
Abstract availability	*Abstract*

Both in arXiv and JHEP, full text is available, as well as abstract.

Download time	*Size of file*
	Network load

The size of the files containing the full text is sometimes available in arXiv Date field of the record. It is not directly available in JHEP records. In both cases, it can be easily obtained as long as the file is available. The network load remains to be known to provide an estimated download time. If a network load is known, the Download time can always be calculated in combination with the size of file. Download time changes over time, and also depends on the user's connection device bandwidth.

Durability	*Anticipated future*
	E-publisher's time policy

Durability can be estimated by the database provider, if once submitted, a document cannot be withdrawn from the database. It is the case for published JHEP documents, where durability is always maximum. For JHEP submitted documents, the e-publisher's policy must be defined: are submitted but not accepted documents withdrawn from the database, or kept with restricted access to editors, or made available to all users? Similarly for arXiv, the policy of the e-publisher can be that once submitted, even the author cannot withdraw a document from the database.

Viewability	*Formats of full text*
Printability	*Formats of full text*

Viewability and printability depend on the formats in which full texts are available, as well as on each particular user's tools. Hence, the values of these two quality features can be improved if the PORTAL makes appropriate viewing and printing tools available for download.

For arXiv documents, formats of full text are sometimes given in the Comments field, and sometimes in optional fields such as Notes and TeX-Type, but it is not given in a well identified specific field. We can think of a simple tool that extracts the information from this field when it is present, but we must also think of adding a special field to be filled up at upload, for future submissions. Generally only one file (hence one format) is available, mostly tex or latex.

For JHEP, files are always available in tex or latex format, with the type given in field Type, as well as in dvi, ps and pdf formats as soon as they are accepted for publication.

Conclusion for availability

Full text availability and Abstract availability are always true in the databases we are concerned with, namely JHEP and arXiv. Nonetheless, in the perspective of opening the tools to other databases, we shall account for these features.

Citability is not always true for arXiv, though true for JHEP documents. A tool that checks the completeness of bibliographic references is needed to evaluate this feature. Submission rules for arXiv may improve this availability. In addition, arXiv users could be asked for their contribution if they are able to provide the full bibliographic reference of a document to which it is not available (see for example "Citeseer Research Index", http://citeseer.nj.nec.com/). This tool, or similar tools, may already exist in the CERN library. In any case, it requires the help of librarians to be designed.

An estimation of the download time feature should be easily made up with a simple tool that computes the size of the available file, and with a tool that estimates the load of the network. It is likely that a rough estimation is enough. The size of the file could even be a sufficient, and more reliable indicator of the download time for users.

Durability may be known with the help of database publishers policy. JHEP and arXiv should be able to provide the information, or to give hints to estimate it if there is no general rule.

Printability and viewability are easy to estimate as formats of full text are available. The missing thing is to extract the format of full text from various arXiv fields. This should be improved in defining a dedicated field filled up at upload time.

5. Authority

Authority can be evaluated directly by a reader of the document, who can use the various subfeatures or give an overall evaluation without going into the details of the subfeatures.

As for a global value, if nothing else is available, it can also be estimated on the basis of *Publication status* (if it has been published).

For arXiv, *Publication status* is sometimes available in Journal-ref or Report-no or Comments or Notes fields, but if it is not specified in one of these fields, it does not mean that the document has not been published.

For JHEP, *Publication status* is always available as the editor will maintain it.

Publication status is a both straightforward and trustful indicator of authority of a paper, which does not need additional evaluation by experts. The problem is that it can only be used for documents that are destinated to publication in some way. For instance it is not applicable for lecture notes.

Reputation of ...

Authority also depends on features that refer to the reputation of something. Reputation can be automatically estimated through a social-oriented tool, or explicitly given by experts.

For a computer-based estimation, the *Citation index* is an important feature, that is not directly available in JHEP nor in arXiv. The issue for these reputation features is that they depend on all existing publications, which in turn are not necessarily present in the considered databases.

They could be obtained from other databases which maintain such features, or built up out of arXiv data, under the hypothesis that arXiv is sufficiently representative of the existing publications. As the full text of documents are available, the first step would be to extract citations from all of the documents, and to deduce the citation index and self-citations. To obtain a more reliable estimation, citations should not be simply counted, but also weighted by the reputation of each citing document.

To conclude, this is a tough work to automatically estimate all reputations mentioned here, and in turn authority. In addition, these estimations may vehicle little authority themselves, as not being reliable.

For human evaluation, an expert is needed, and the following issues arise:

1. size of set of evaluated items (how many items will experts have to assess reputation for)
2. maintainance of set of evaluated items (how often the set of evaluated items evolves)
3. maintainance of reputations (how often reputations shall be updated)
4. control of possible biases (experts may be biased in their evaluation)

We are going to consider one by one each reputation feature, with respect to each of these issues.

Reputation of author..*Author*
....*Citation index*

Author is always available. Authors reputations could be assessed by experts, but there are many of them, and there is an ever growing number of authors. Moreover, a new author may quickly gain a good reputation. Hence experts will have a tough work in maintaining the reputation scores. Automatic processes should help them to do it, not only to make it feasible, but also to avoid biases.

Examples of helping tools are:
Track arising authors (to have experts assess only important authors)
Control biases via cross-checking scores coming from several experts
Check scores with citation-index and other available reputation

Reputation of organization..............................*Organization of author*

Organization of the author is not always available, but it can be an interesting shortcut to estimate the *Reputation of author* feature, as there are much less organizations than authors, they are more stable, and their reputations is not changing fast. This makes it easier for experts to assess the reputation of organizations.

The question of being aware of new organizations, and of helping experts to evolve their scores remains. Automatic processes should be designed to
Track new organizations
Control biases via cross-checking scores coming from several experts
Check scores with citation-index and other available reputation

Reputation of publisher .... *Publisher*

The remarks given for reputation of organizations hold for publishers.

Reputation of editor.. *Editor*

For reputation of editors, the issue is slightly different as there may be a greater number of editors, especially as editorial boards are often made of several persons (experts), and are likely to be gathered and modified frequently. Nonetheless, the issue seems smoother than for authors, as editorial board members are generally well known persons.

Reputation of journal or conference................ *Journal or conference*

For reputation of journal or conference, the remarks are similar to those for reputation of organization and publisher, as the number is smaller.

Conclusion for authority
As a conclusion for Authority, the most feasible approach is to use *Publication status*, in combination with expert evaluation of *Reputations of...*.
When *Publication status* is "published in refereed publication", then reputations of publisher, editors, and journal or conference can be used to refine the score.
When *Publication status* is different from "published in refereed publication", then *Reputation of organization* and *Reputation of author* must be used to refine the judgment.

An important point is to try and improve the availability of the *Publication status* feature in arXiv records.

6. Popularity

Popularity is another social-related quality feature, which has a non-expert taste, as compared to authority. It is meant to reflect the actual use of documents in the community, rather than the opinion that experts could have on them.

Citation index
Downloads
Recommandations

A good indicator of popularity is the number of downloads of a paper. These are available for arXiv, as logs of all http connections are archived. They should be available for JHEP too (to be confirmed).

Recommandations could only be available from a social tool which tracks informal exchanges between peers in the community (in forums, in favorite links in Web pages). They will not be considered here.

Finally, citation index also gives hints on documents usage. But as we said for authority, they are not easily obtained in this context.

7. Conclusion

To conclude, many simple tools to extract information from fields that contain free-text information are needed to improve estimation of missing document features. As an alternate and complementary solution, suggestions of evolution in the upload procedure are given.

More complicated are the tools needed to estimate social aspects (essentially reputations for authority), as no citation index is available. Nonetheless, downloads can give a good indicator of popularity.

To estimate intended audience, readability, and scientific quality, automatic methods are uinlikely to be successful, although the feasibility of heuristics can be discussed with experts for some of the features. Complementarily, these features are typically those used for peer review. The relevance of the various quality features mentioned there should be validated by experts (editors, scientists used to reviewing, etc.).

For some features, the participation of simple readers can bring a lot, like for completion of incomplete bibliographic references, or intended audience features.