Rating repositories

Handy shortlink to this document: http://bit.ly/rating-repositories

What should we be rating repositories on?

Repositories should be rated on compliance with relevant independent third party standards. Different repositories have different purposes and some standards that are relevant to some repositories may not be relevant to other repositories.

How should we be rating repositories?

Repositories should be rated as far as possible in ways that:
1) Can be automatically detected / validated
2) Can be used to suggest steps for improvement
3) We should not be afraid to include goals which are not currently easily attainable but which are widely accepted as desirable.
4) We should not be afraid to include goals which are mutually incompatible
5) Lead to cooperation and improvement of the community

Which specific measures should be used?

Below are some ideas. Not all of these will be appropriate for every repository, but all of them are worth consideration by repository admins. Not all of them may turn out to be reliably testable-for, but again they're worth consideration by repository admins.

Here we list specific script-testable things based on specified standards / recommendations / guidelines:

From http://www.w3.org/Provider/Style/URI.html
1. ...
2. To enable persistence, URIs are assigned based on a regular scheme (URIs have low entropy with respect to each other)

From http://www.w3.org/TR/cooluris/
1. ?

From http://validator.w3.org/feed/ (RSS)
1. at least one well-formed feed is produced
2. at least one HTML page correctly advertises the feed in the header
3. at least one HTML page correctly advertises the feed in the body
4. the feed is correctly UTF-8 encoded.
5. the feed uses recommended dc fields

From http://www.openarchives.org/OAI/openarchivesprotocol.html
and http://www.openarchives.org/OAI/2.0/guidelines.htm ,
http://www.openarchives.org/OAI/2.0/guidelines-repository.htm
1. repository is harvestable via OAI-PMH
2. OAI-PMH link is advertised
3. standard OAI verbs (Identify, ListSets, ListMetadataFormats, ListRecords, ListIdentifers, GetRecord) generate a valid OAI response
4. declaration of support for "deletion mode" in Identify response (either 'no', 'transient', or 'persistent') and compliance with this declaration
5. support for 'persistent' deletion mode, and compliance with this declaration
6. simple Dublin Core metadata format supported, and recommended fields exposed
7. "marc21" metadataPrefix supported
8. "oai_marc" metadataPrefix supported
9. "oai_dc" metadataPrefix supported
10. root metadata node in response contains complete and valid XML namespace(s)
11. the response is correctly UTF-8 encoded

From http://dx.doi.org/
1. DOI is used
2. DOI is advertised
3. DOIs can be resolved correctly
4. DOIs can be resolved over pure-HTTPS

From http://www.handle.net/ (greyed out, may not be best-practise any more?)
1. if advertised, handles resolve correctly
2. handles are used
3. handles are advertised
4. handles can be resolved over pure HTTPS

...

From http://tools.ietf.org/html/rfc2818
1. The repository is accessible via HTTPS
2. The repository login isn't over HTTP
3. SSL certificate isn't self-signed
4. Reuse tool at
   https://www.ssllabs.com/ssltest/analyze.html?d=exams.victoria.ac.nz ?

From http://www.dnssec.net/

1. DNSSEC is used

From http://ec.europa.eu/ipg/basics/legal/cookies/index_en.htm
1. The site uses no cookies
2. The site uses explicit cookie notification
3. No known-abusive privacy-invasive sites are accessed
4. Meets EU cookies policy

From http://dublincore.org/documents/dces/
1. dc.title used
2. dc.creator used
3. dc.date used
4. dc.description used
5. dc.identifier used
1. at least some items have dc.rights set
2. at least some items have dc.rights.uri set
3. at least some items have a dc.rights URI set to outside the domain of the repository.
4. at least some items have a dc.rights URI set to http://*.creativecommons.*/
5. all items have a dc.rights URI
6. …

From http://www.unicode.org/reports/tr10/
1. Unicode searching works (pick a substring from an item, change case or accents, check that the item is still found)

From http://www.robotstxt.org/orig.html and http://www.sitemaps.org/protocol.html
1. Robots txt is present
2. Robots txt has a sitemap or htmlmap
3. Sitemap or htmlmap is up to date
4. ...
   Demonstrator for such a thing implemented on
   http://dspacecheck-atmire.rhcloud.com/

From https://www.ietf.org/rfc/rfc1952.txt (compressed HTTP)
1. Server support compression for efficient downloading of large files

From https://www.ietf.org/rfc/rfc2046.txt (mimetype)
1. Mimetypes are correctly used to identify content accessible for download

From: http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=38920
1. PDFs being served use the http://www.aiim.org/pdfa/ns/id/ namespace (claim to be PDF/A)

2. PDFs being served claim to be PDF/A-1
3. PDFs being served claim to be PDF/A-2
4. PDFs being served claim to be PDF/A-3
5. PDFs being served validate as meeting their claims
6. PDF document metadata matches item-level document metadata (dc.title, etc)
7. PDF document metadata links back to the repository URL or identifier
8.

From http://www.iso.org/iso/catalogue_detail?csnumber=51463 (Microsoft Office XML docs)
1. Documents are not encrypted
2. Documents are not locked
3. PDF document metadata matches item-level document metadata (dc.title, etc)
4.

From http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=43485
(Open Office XML docs)
1. Documents are not encrypted
2. Documents are not locked
3. PDF document metadata matches item-level document metadata (dc.title, etc)
4.

From http://www.w3.org/DesignIssues/LinkedData.html
1. URIs are HTTP or HTTPS
2. Content negotiation is supported
3. Some form of RDF via content negotiation is supported
4. JSON-LD via content negotiation is supported
5. dc:creator, dc:identifier (and others?) provide Linked Data URIs (1-3 above)
6. There are URIs elsewhere in this domain
7. There are URIs that are not from this domain
8. There are URIs from other repositories
9. The repository is searchable via some SPARQL endpoint
10. The SPARQL endpoint is advertised
11. The sitemap has a sc:dataset according to the Semantic Sitemap extension (http://www.w3.org/wiki/SemanticSitemap)
12.

From http://www.w3.org/TR/rdfa-syntax/
1. RDFa is embedded in the HTML
2.

From https://en.wikipedia.org/wiki/Citation
1. APA citation for the item is given
2. MLA citation for the item is given
3. Chicago citation for the item is given

4. RIS citation for the item is given
5. Bibtex citation for the item is given
6. Wikipedia citation for the item is given
7. ...

Here we list specific script-testable interconnects

To http://repec.org/
1. At least one collection has valid redif; see validator at http://econpapers.repec.org/scripts/redifcheck.pl
2. Items deposited in at least one series are more than X days ago are transferred to a series in repec.

To http://oaister.worldcat.org/
1. Items deposited more than ? months ago are locatable via OAIster catalogue
2. OAIster catalogue features no items withdrawn from the repository
3. Metadata of harvested item in OAIster matches basic DC record of item in repository

To http://scholar.google.co.nz/
1. Items deposited more than X months ago are locatable via google scholar.

To other OAI -based search engines:
1. http://base-search.net/
2. http://academic.research.microsoft.com (comes in via OAIster)
3. http://oanet.cms.hu-berlin.de/oansearch/ (currently only in german)

OAI validation tools/frameworks
1. http://oanet.cms.hu-berlin.de/validator/pages/validation_dini.xhtml (currently only in german)
2. https://www.openaire.eu/validator
3. http://validator.oaipmh.com/
4. http://re.cs.uct.ac.za/
5. http://oval.base-search.net/ codebase: https://github.com/pietsch/oval
6.

Example rankings out there

http://repositoryranking.org/