A Checklist-Based Approach for Quality Assessment of Scientific Information

Jun Zhao, Graham Klyne
Department of Zoology, University of Oxford,
Oxford, OX1 3PS, UK
jun.zhao, graham.klyne@zoo.ox.ac.uk

Matthew Gamble, Carole Goble
Computer Science, University of Manchester, Manchester,
M13 9PL, UK
m.gamble@cs.man.ac.uk,
carole.goble@manchester.ac.uk

Workflow 4Ever

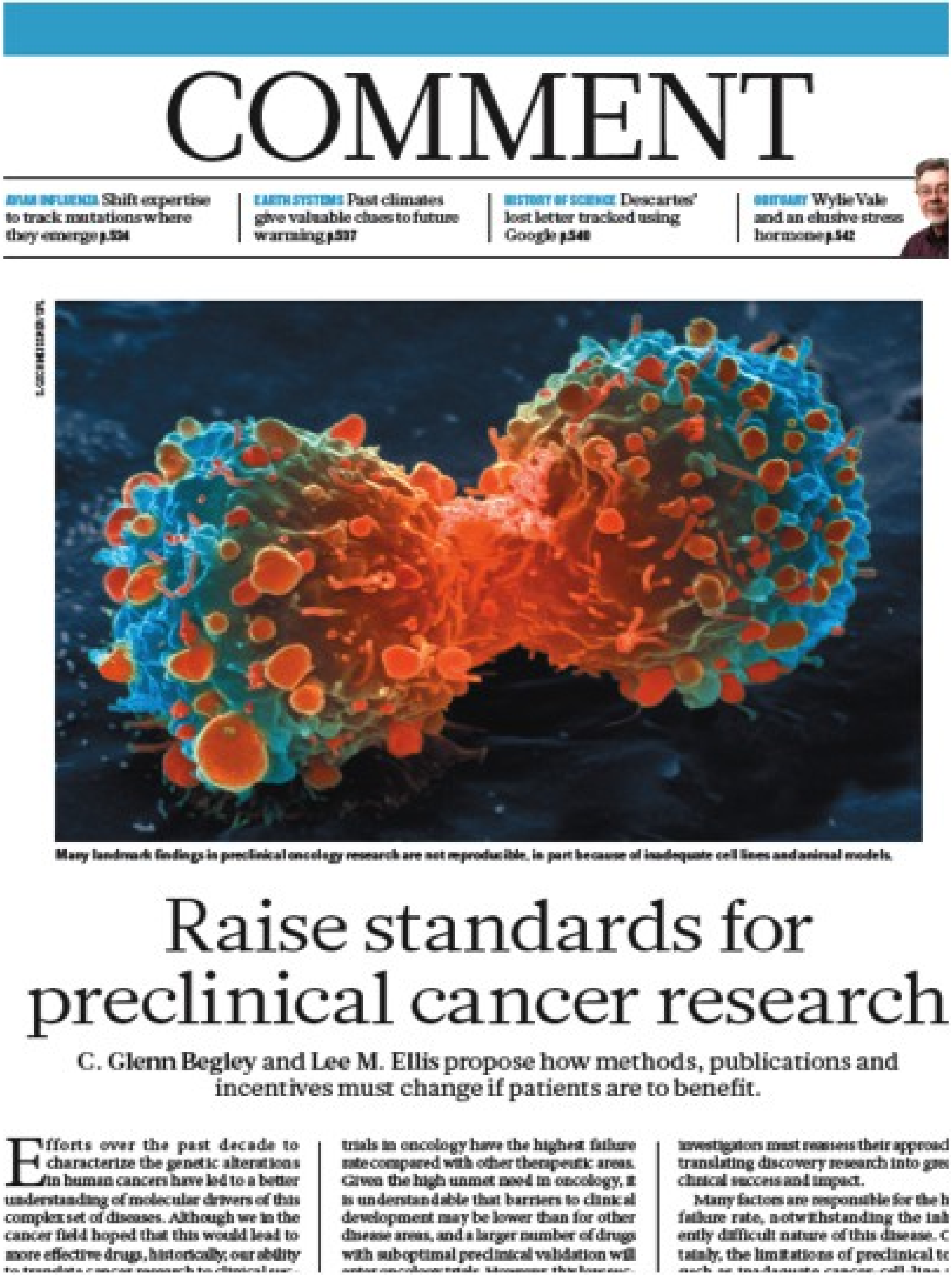
47 of 53 “landmark” publications could not be replicated

“Unquestionably, a significant contributor to failure in oncology trials is the quality of published preclinical data.”

“The scientific process demands the highest standards of quality, ethics and rigour.”

http://www.nature.com/nature/journal/v483/n7391/full/483531a.html

Adapted from: Carole Goble



COMMENT

AVIAN INFLUENZA Shift expertise to track mutations where they emerge p.534

EARTH SYSTEMS Past climates give valuable clues to future warming p.537

HISTORY OF SCIENCE Descartes' lost letter tracked using Google p.540

OBITUARY Wylie Vale and an elusive stress hormone p.542

Many landmark findings in preclinical oncology research are not reproducible, in part because of inadequate cell lines and animal models.

Raise standards for preclinical cancer research

C. Glenn Begley and Lee M. Ellis propose how methods, publications and incentives must change if patients are to benefit.

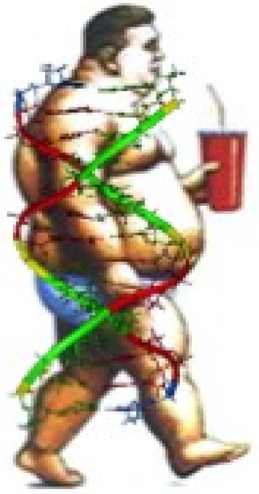
Efforts over the past decade to characterize the genetic alterations in human cancers have led to a better understanding of molecular drivers of this complex set of diseases. Although we in the cancer field hoped that this would lead to more effective drugs, historically, our ability

trials in oncology have the highest failure rate compared with other therapeutic areas. Given the high unmet need in oncology, it is understandable that barriers to clinical development may be lower than for other disease areas, and a larger number of drugs with suboptimal preclinical validation will

investigators must reassess their approac translating discovery research into gre clinical success and impact.

Many factors are responsible for the h failure rate, notwithstanding the inh ently difficult nature of this disease. C tainly, the limitations of preclinical t

Validating *in silico* research



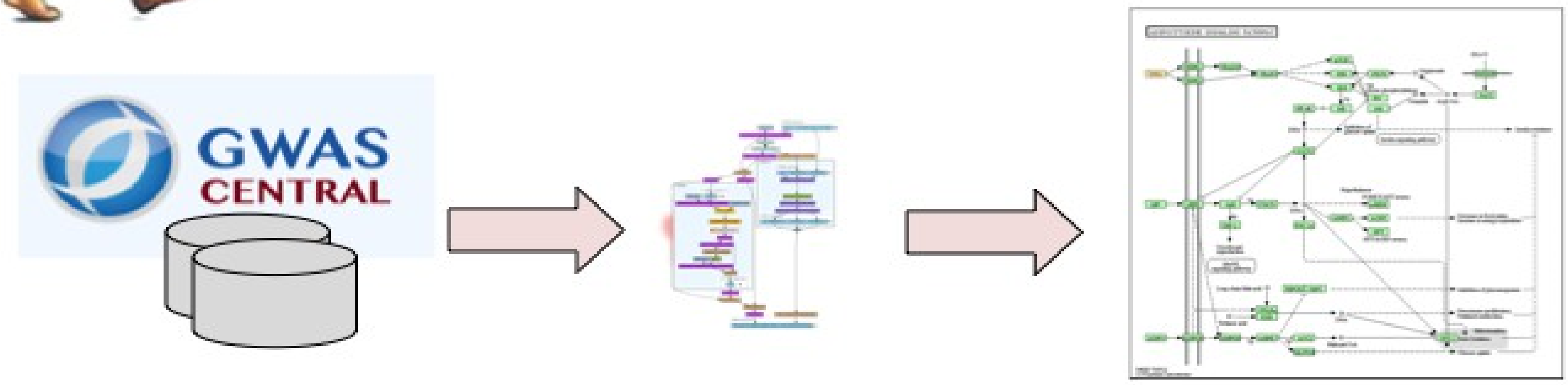
Research Question

Genome Wide Association Studies (GWAS)

In 1000+ people: which gene mutations are associated with metabolic syndrome, and why?

Hypothesis

Genes involved in inflammation pathways are involved in the onset of metabolic syndrome.



Download data
- External DB
- Existing Knowledge

Workflow:
Which biological pathways explain the associations?

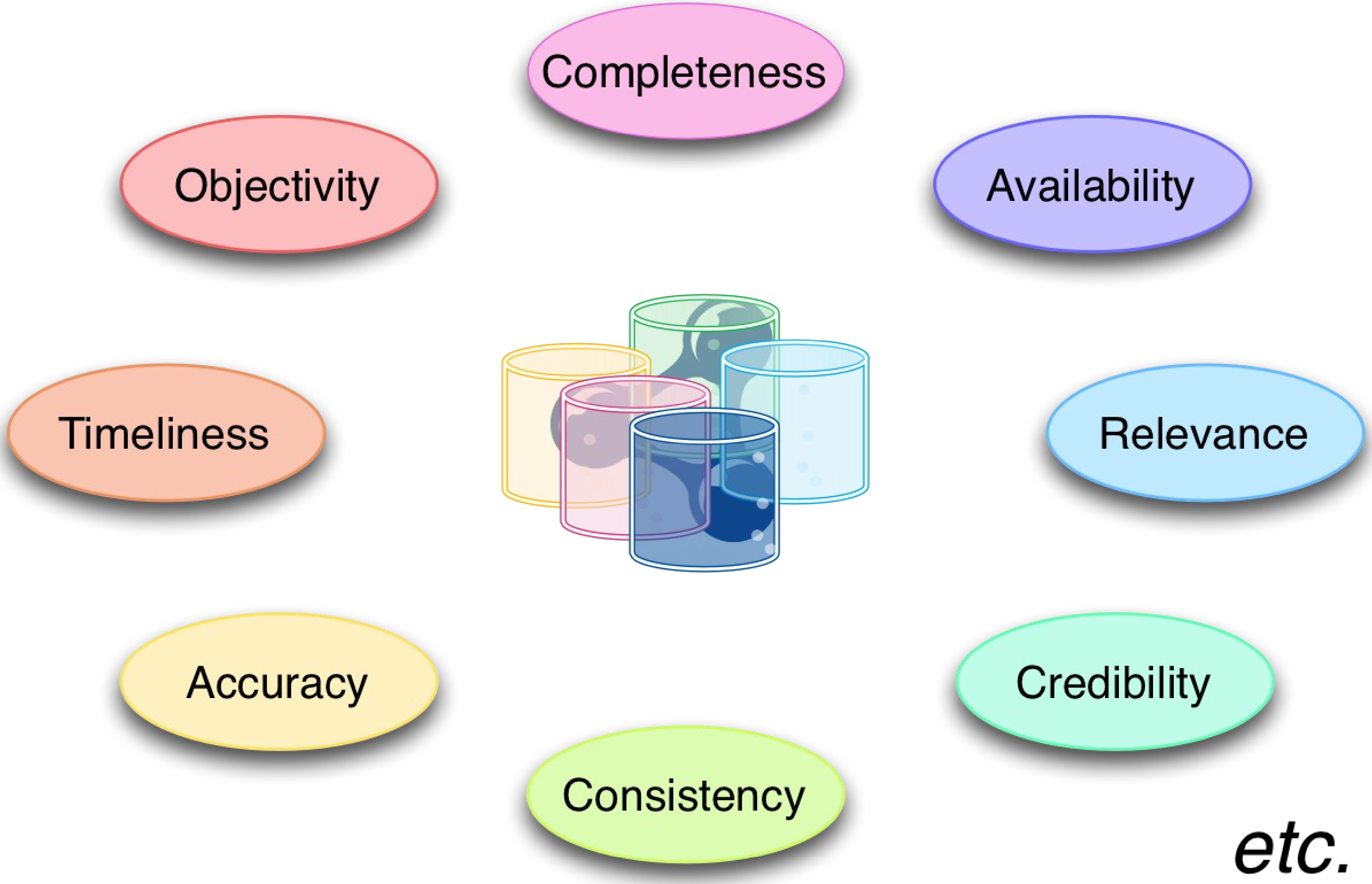
Interpret results
(Interaction pathways in the cell)

Kristina Hettne

The Goal of Our Work

We aim to provide tools that assess the ***quality of information*** (***data*** and ***computational artifacts***) used and generated by researchers

Background:
Linked Data Quality



Completeness

Objectivity

Availability

Timeliness

Relevance

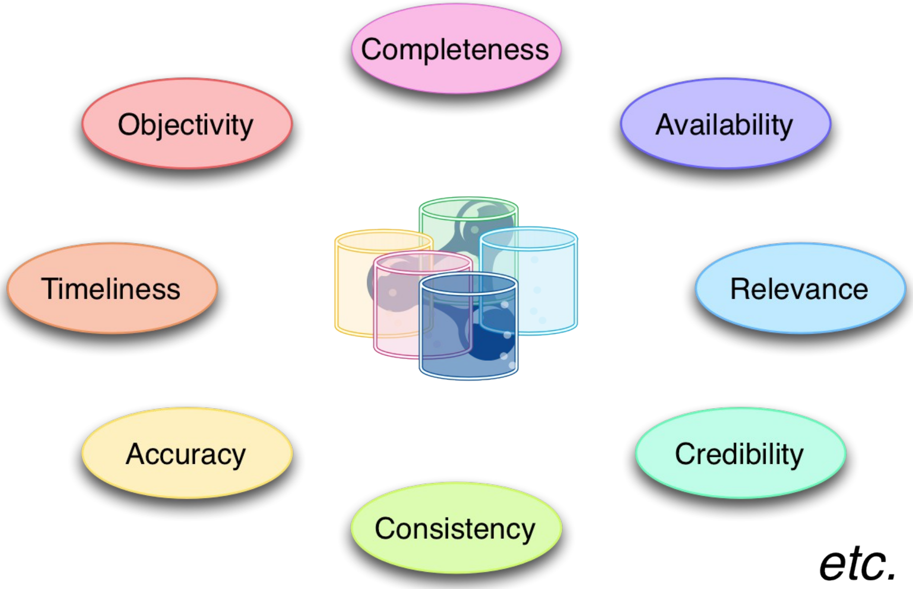
Accuracy

Credibility

Consistency

etc.

Information quality



etc.

http://upload.wikimedia.org/wikipedia/commons/3/32/Blind_men_and_elephant.jpg

For these elephant *users*...

http://www.flickr.com/photos/leosoueu/7121406603/

For these elephant *users*…

… “can this elephant get me across this river?”

http://www.flickr.com/photos/leosoueu/7121406603/

Our approach to information quality assessment

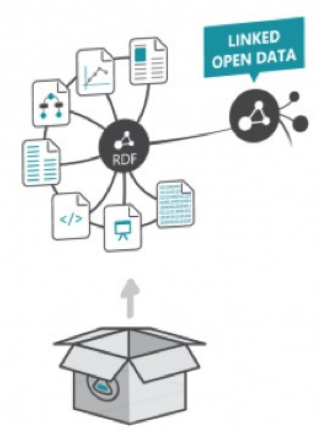
We see complementary approaches to quality:

- An analysis of what constitutes quality, and
- Users' concern about fitness for use

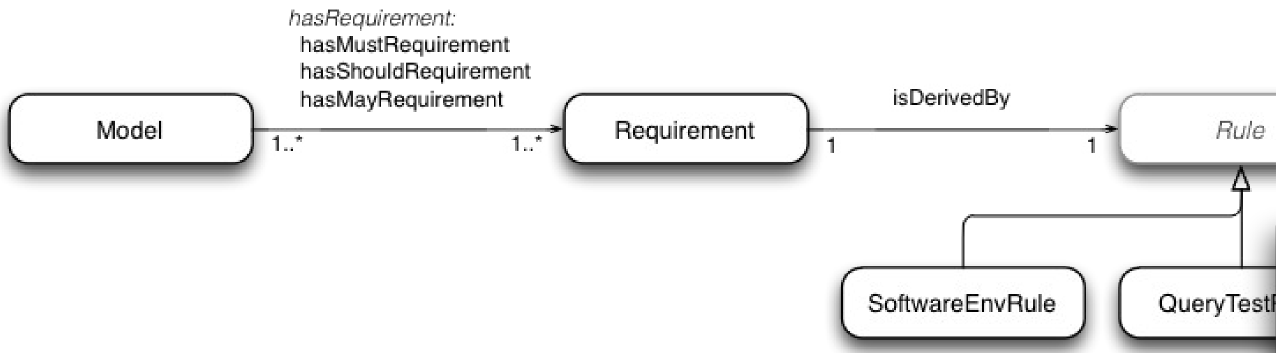
The work presented here is focused on the latter of these concerns

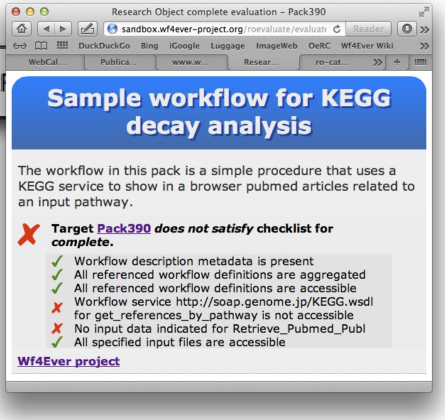
Outline

Research Objects: Context for evaluation



Checklists and the “Minim” model





Sample workflow for KEGG decay analysis

The workflow in this pack is a simple procedure that uses a KEGG service to show in a browser pubmed articles related to an input pathway.

✗ **Target Pack390 *does not satisfy* checklist for *complete*.**

- ✓ Workflow description metadata is present
- ✓ All referenced workflow definitions are aggregated
- ✓ All referenced workflow definitions are accessible
- ✗ Workflow service http://soap.genome.jp/KEGG.wsdl for get_references_by_pathway is not accessible
- ✗ No input data indicated for Retrieve_Pubmed_Publ
- ✓ All specified input files are accessible

Wf4Ever project

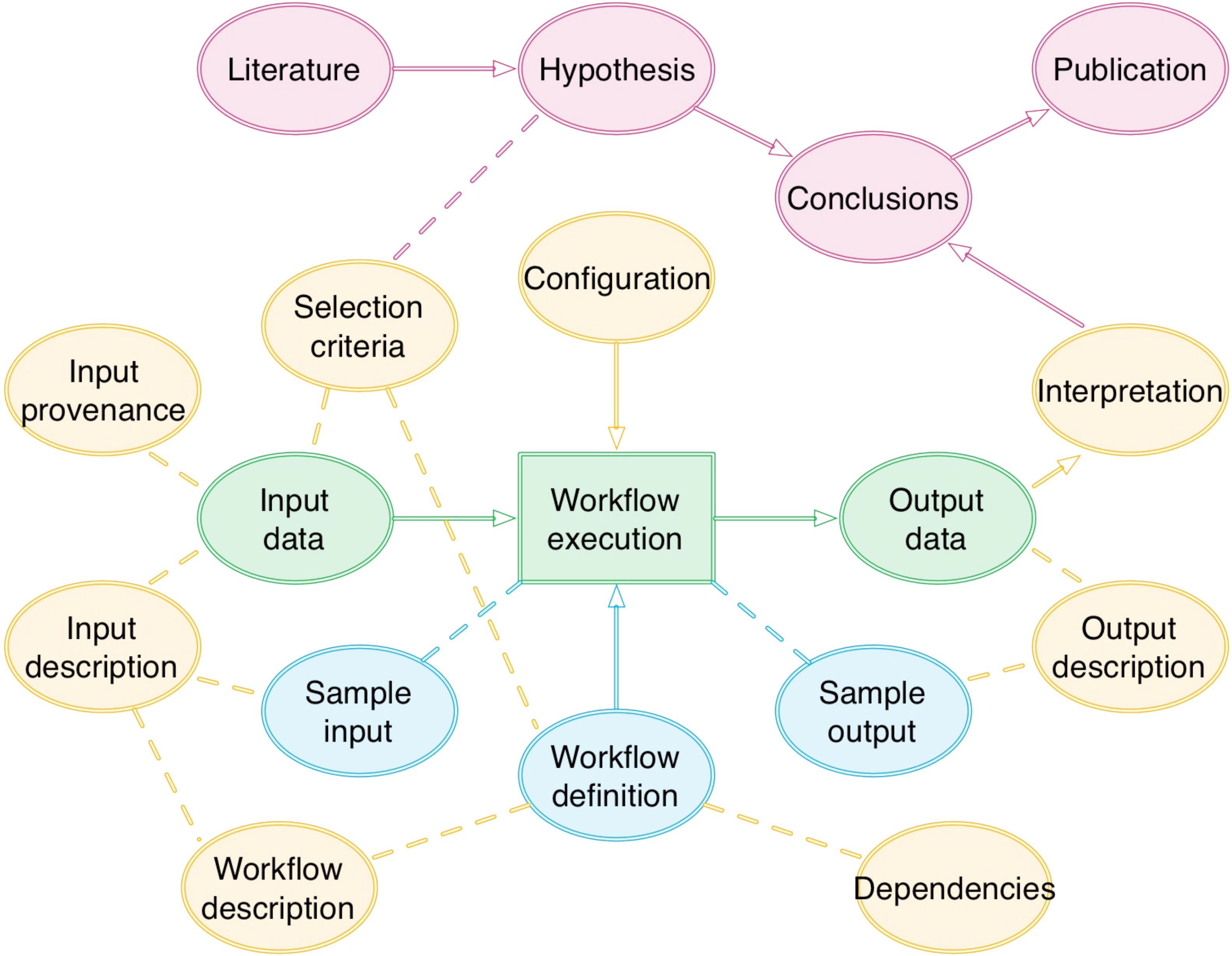
Evaluation and reflections

Continuing work and concluding remarks

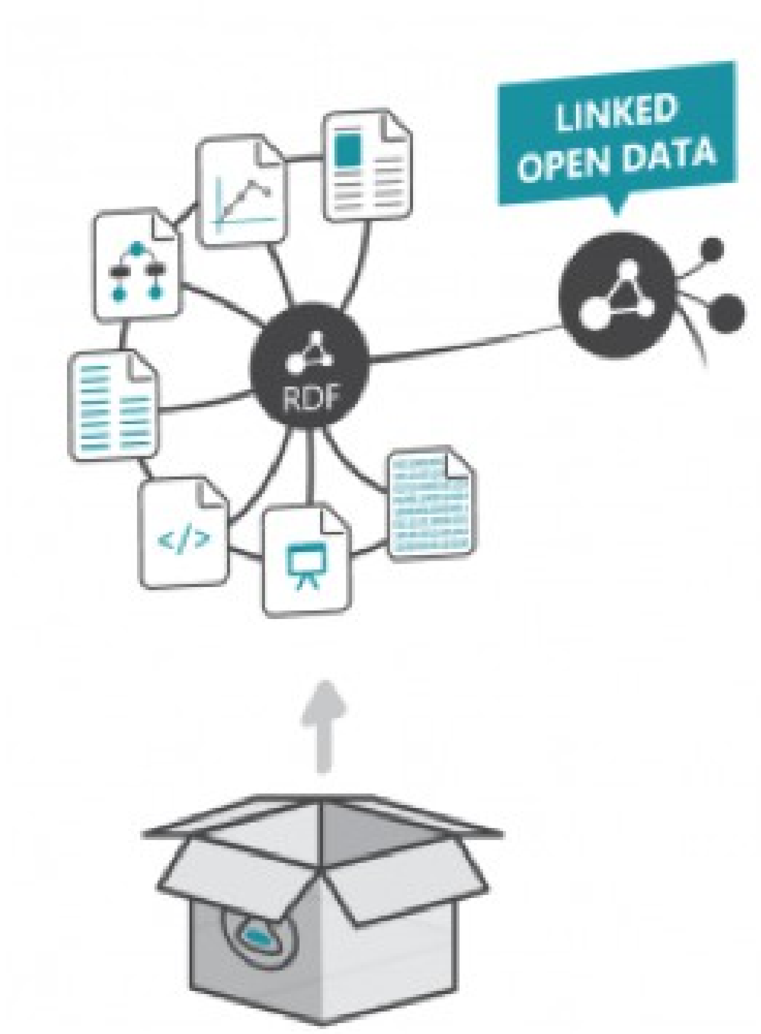
Research Objects

Context for information quality evaluation

Context for evaluation of scientific information quality



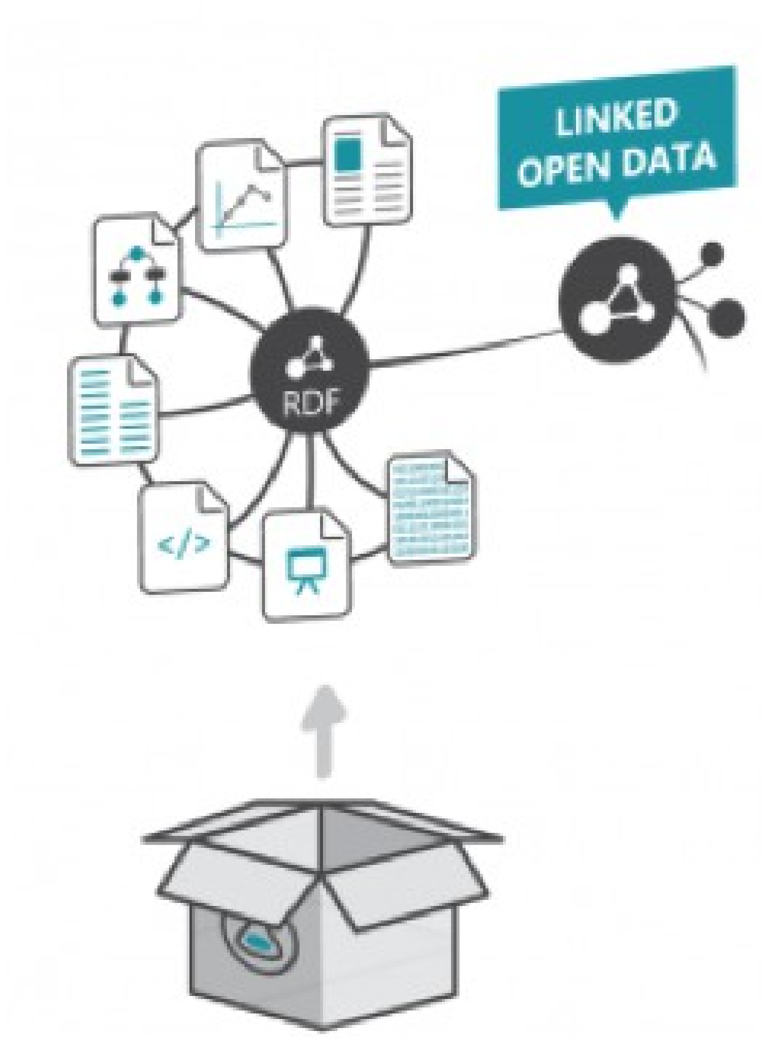
Context for evaluation: Research Objects



- **Data** used or results produced in an experiment study
- **Methods** employed to produce and analyse that data
- **Provenance** and **setting information** about the experiments
- **People** involved in the investigation
- **Annotations** about these resources, that are essential to the understanding and interpretation of the scientific outcomes captured in a research object.

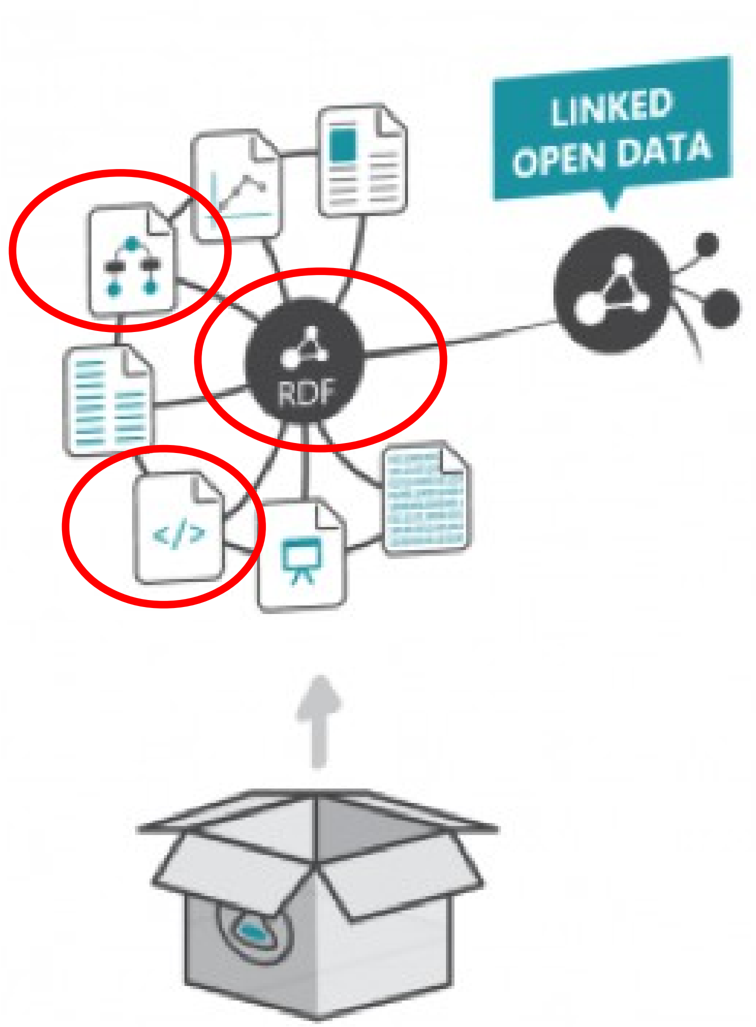
Jun Zhao

Context for evaluation: Research Objects



- **Annotations** about these resources, that are essential to the understanding and interpretation of the scientific outcomes captured in a research object.

Quality evaluation uses RO annotations



In our RO implementation, **annotations** are presented as RDF

E.g., annotations for:

- types of aggregated resources (data, workflow, result, hypothesis, etc.)
- workflow element descriptions
- workflow run provenance traces
- … and more

These annotations are **merged** into a single RDF graph to provide a starting point for the evaluation process

Fitness-for-use of a Research Object

Our approach is intended to address suitability of RO content, and particularly replicability and reproducibility of results, such as:

- *Can I trust the conclusion of the experiment described in this RO?*
- *Do the workflows used in this RO still work?*
- *Is the investigation described in this RO ready for publication?*
- *Can I re-purpose the workflow used in this RO for my own experiment?*

Some scenarios

Workflow decay detection

- *does the workflow used by this RO still work?*
- *are the external resources used (still) accessible?*

DBPedia “ChemBox” data completeness

- *does chemical data extracted from Wikipedia info boxes meet community expectations for a full description of a chemical?*

Workflow best practices

- *has the workflow contained in an RO been supported by good development practices?*

Checklists and the "Minim" model

Our chosen tool: checklists for ROs

Checklists are a widely used, well-understood tool for quality management

A checklist is simply a list of tests; *e.g.*

- does the RO contain an experimental hypothesis?
- does the RO contain a workflow?
- is the executable workflow description accessible?
- do workflows in the RO have defined inputs?
- are the workflow input files accessible?
- are the workflow web services used accessible?
- *etc.*

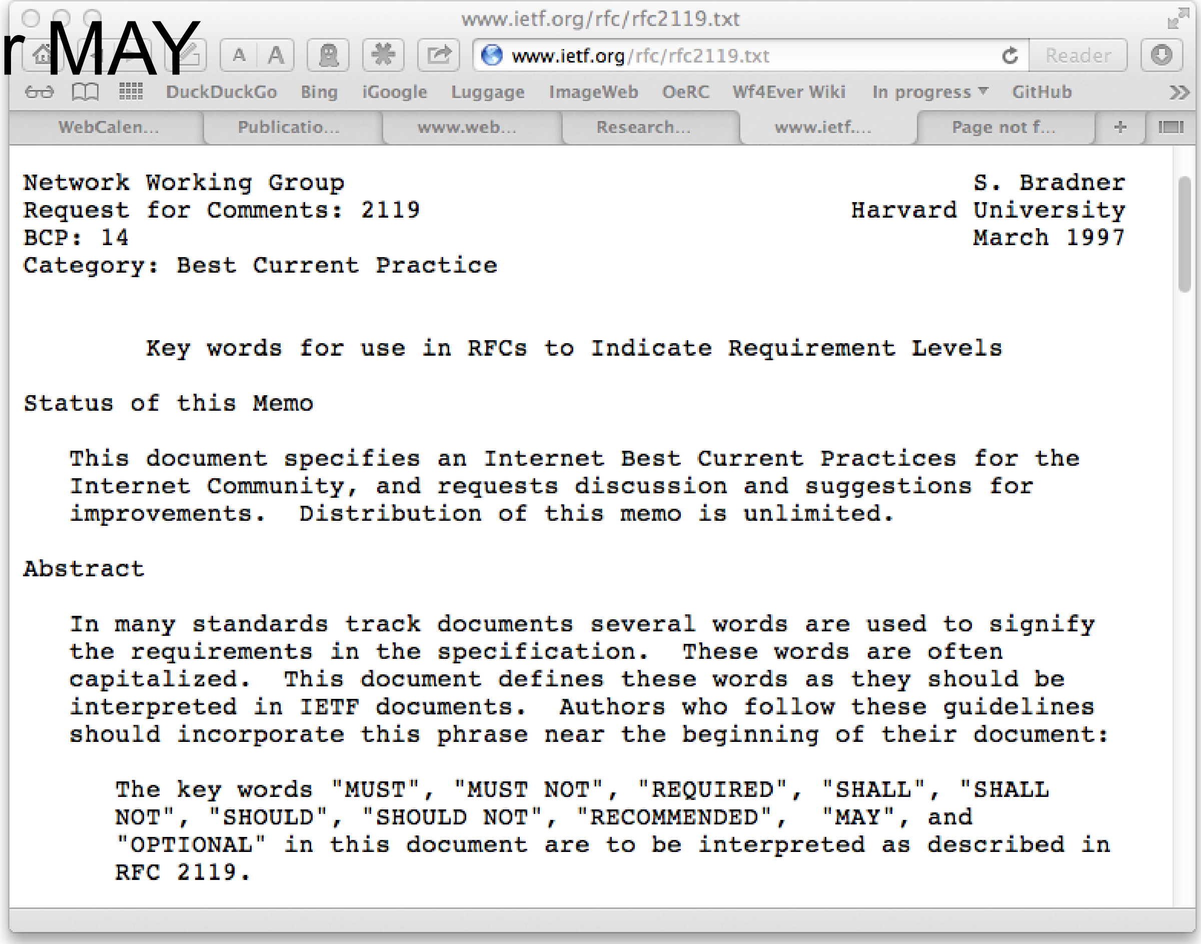
Requirement levels

Each checklist test has a requirement level:

- MUST, SHOULD or MAY
- from RFC 2119

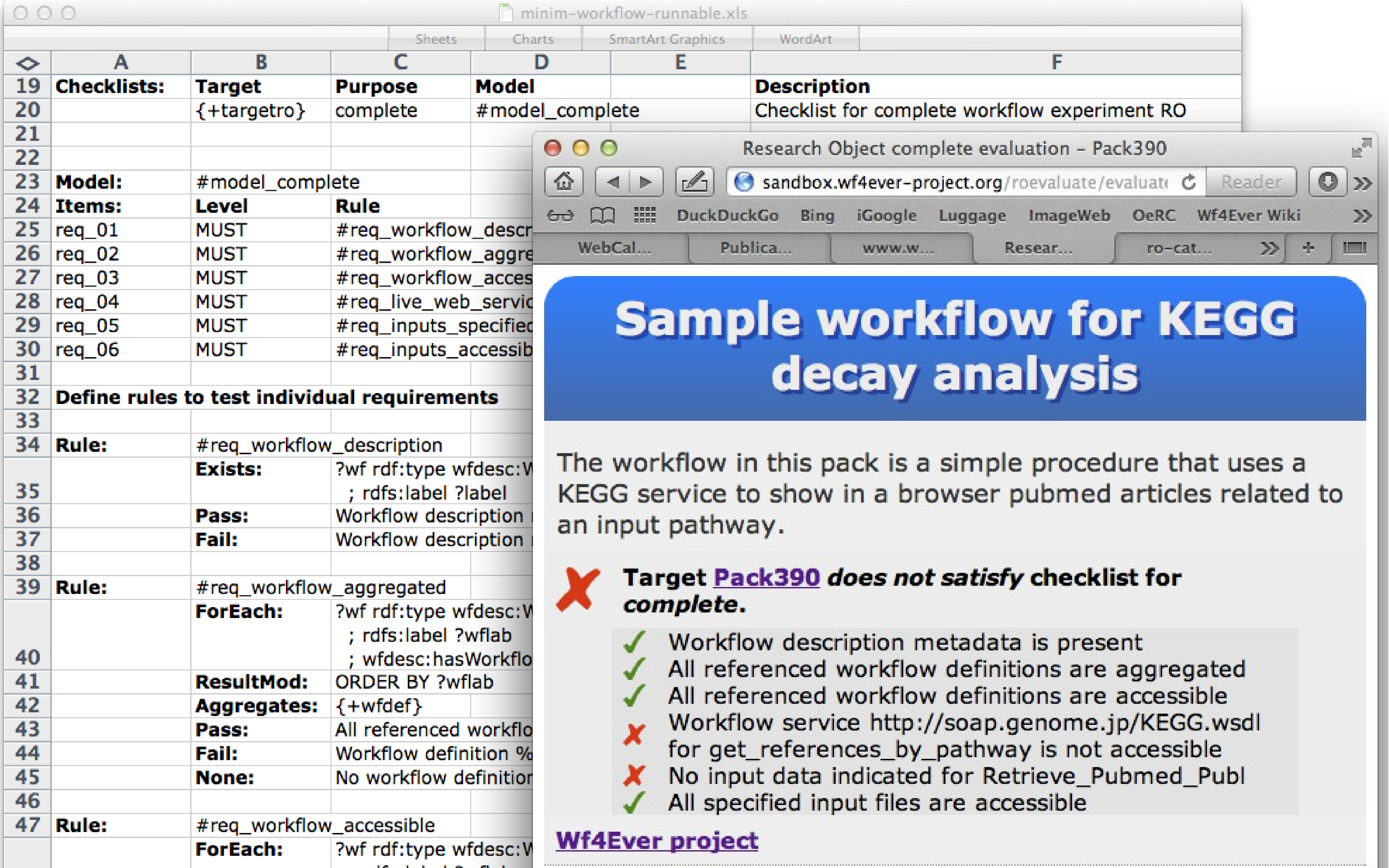
The overall checklist result reflects the level of satisfied and unsatisfied requirements:

e.g. a failed MUST is more serious than a failed SHOULD



```
Network Working Group                                         S. Bradner
Request for Comments: 2119                            Harvard University
BCP: 14                                                       March 1997
Category: Best Current Practice


        Key words for use in RFCs to Indicate Requirement Levels

Status of this Memo

   This document specifies an Internet Best Current Practices for the
   Internet Community, and requests discussion and suggestions for
   improvements.  Distribution of this memo is unlimited.

Abstract

   In many standards track documents several words are used to signify
   the requirements in the specification.  These words are often
   capitalized.  This document defines these words as they should be
   interpreted in IETF documents.  Authors who follow these guidelines
   should incorporate this phrase near the beginning of their document:

      The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL
      NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED",  "MAY", and
      "OPTIONAL" in this document are to be interpreted as described in
      RFC 2119.
```

Runnable workflow checklist



	A	B	C	D	E	F
19	**Checklists:**	**Target**	**Purpose**	**Model**		**Description**
20		{+targetro}	complete	#model_complete		Checklist for complete workflow experiment RO
21						
22						
23	**Model:**	#model_complete				
24	**Items:**	**Level**	**Rule**			
25	req_01	MUST	#req_workflow_descr			
26	req_02	MUST	#req_workflow_aggre			
27	req_03	MUST	#req_workflow_acces			
28	req_04	MUST	#req_live_web_servic			
29	req_05	MUST	#req_inputs_specified			
30	req_06	MUST	#req_inputs_accessib			
31						
32	**Define rules to test individual requirements**					
33						
34	**Rule:**	#req_workflow_description				
35		**Exists:**	?wf rdf:type wfdesc:W ; rdfs:label ?label			
36		**Pass:**	Workflow description			
37		**Fail:**	Workflow description			
38						
39	**Rule:**	#req_workflow_aggregated				
40		**ForEach:**	?wf rdf:type wfdesc:W ; rdfs:label ?wflab ; wfdesc:hasWorkflo			
41		**ResultMod:**	ORDER BY ?wflab			
42		**Aggregates:**	{+wfdef}			
43		**Pass:**	All referenced workflo			
44		**Fail:**	Workflow definition %			
45		**None:**	No workflow definitior			
46						
47	**Rule:**	#req_workflow_accessible				
		ForEach:	?wf rdf:type wfdesc:W			

Sample workflow for KEGG decay analysis

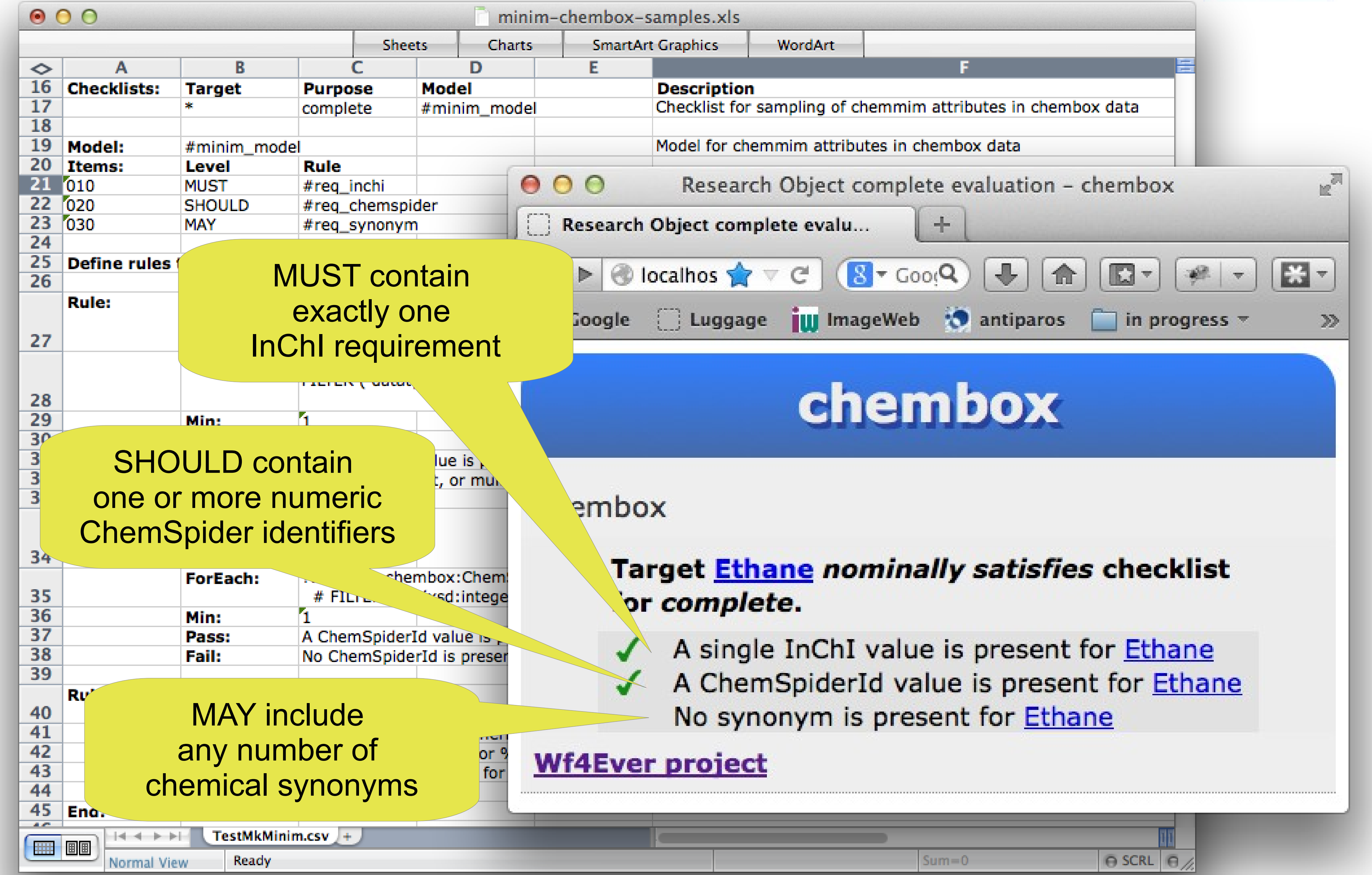
The workflow in this pack is a simple procedure that uses a KEGG service to show in a browser pubmed articles related to an input pathway.

✗ **Target Pack390 *does not satisfy* checklist for *complete*.**

- ✓ Workflow description metadata is present
- ✓ All referenced workflow definitions are aggregated
- ✓ All referenced workflow definitions are accessible
- ✗ Workflow service http://soap.genome.jp/KEGG.wsdl for get_references_by_pathway is not accessible
- ✗ No input data indicated for Retrieve_Pubmed_Publ
- ✓ All specified input files are accessible

Wf4Ever project

Chemical data checklist



	A	B	C	D	E	F
16	**Checklists:**	**Target**	**Purpose**	**Model**		**Description**
17		*	complete	#minim_model		Checklist for sampling of chemmim attributes in chembox data
18						
19	**Model:**	#minim_model				Model for chemmim attributes in chembox data
20	**Items:**	**Level**	**Rule**			
21	010	MUST	#req_inchi			
22	020	SHOULD	#req_chemspider			
23	030	MAY	#req_synonym			
24						
25	**Define rules f**					
27	**Rule:**					
28			FILTER (datat			
29		**Min:**	1			
31				lue is		
32				t, or mu		
35		**ForEach:**	chembox:Chem # FILTE xsd:intege			
36		**Min:**	1			
37		**Pass:**	A ChemSpiderId value is p			
38		**Fail:**	No ChemSpiderId is preser			
40	**Ru**					
42				or %		
43				for		
45	**End**					

Research Object complete evaluation – chembox
Research Object complete evalu...
localhos
Goo
Google
Luggage
ImageWeb
antiparos
in progress

chembox

embox

Target Ethane *nominally satisfies* checklist for *complete*.

✓ A single InChI value is present for Ethane

✓ A ChemSpiderId value is present for Ethane

No synonym is present for Ethane

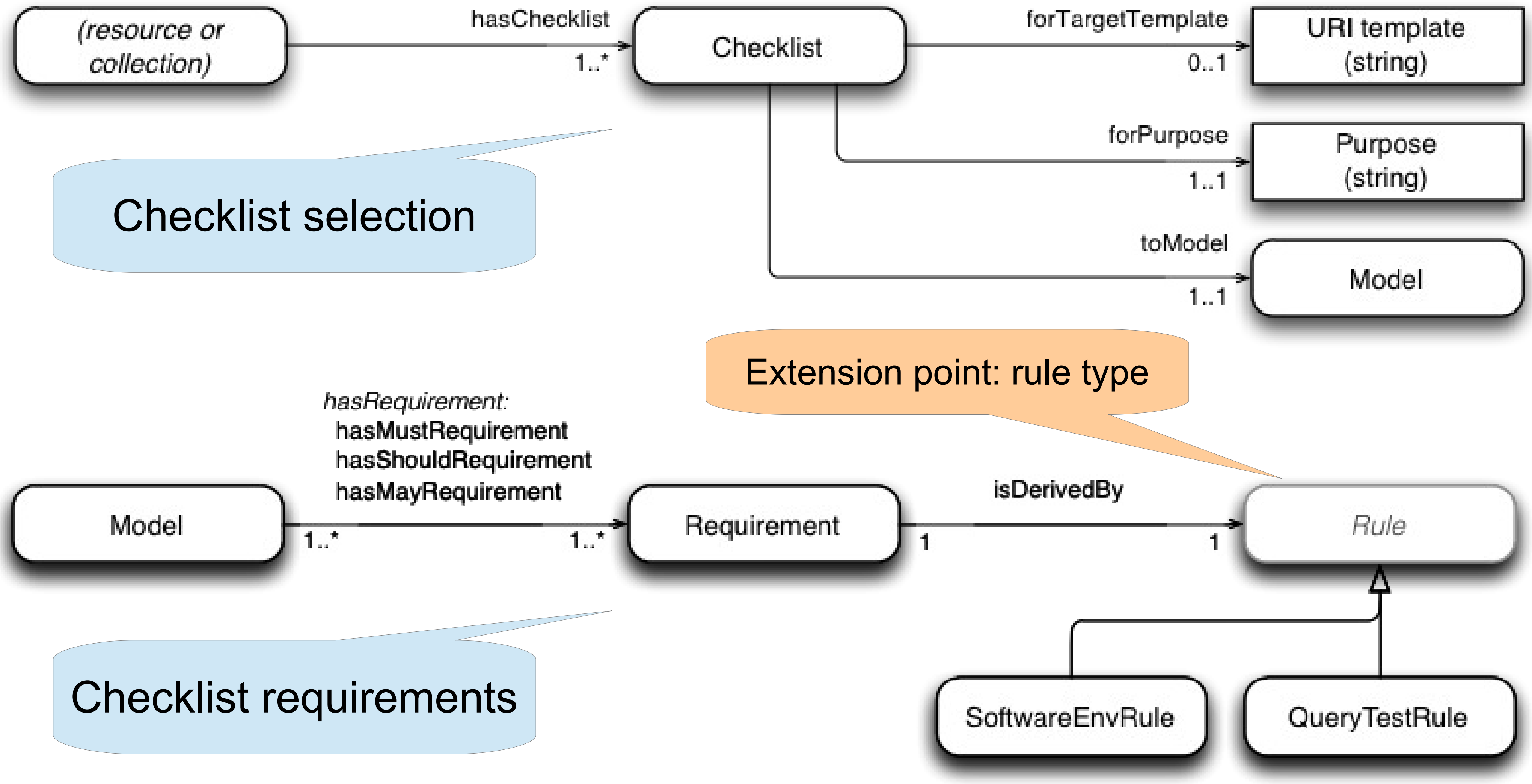
Wf4Ever project

MUST contain exactly one InChI requirement

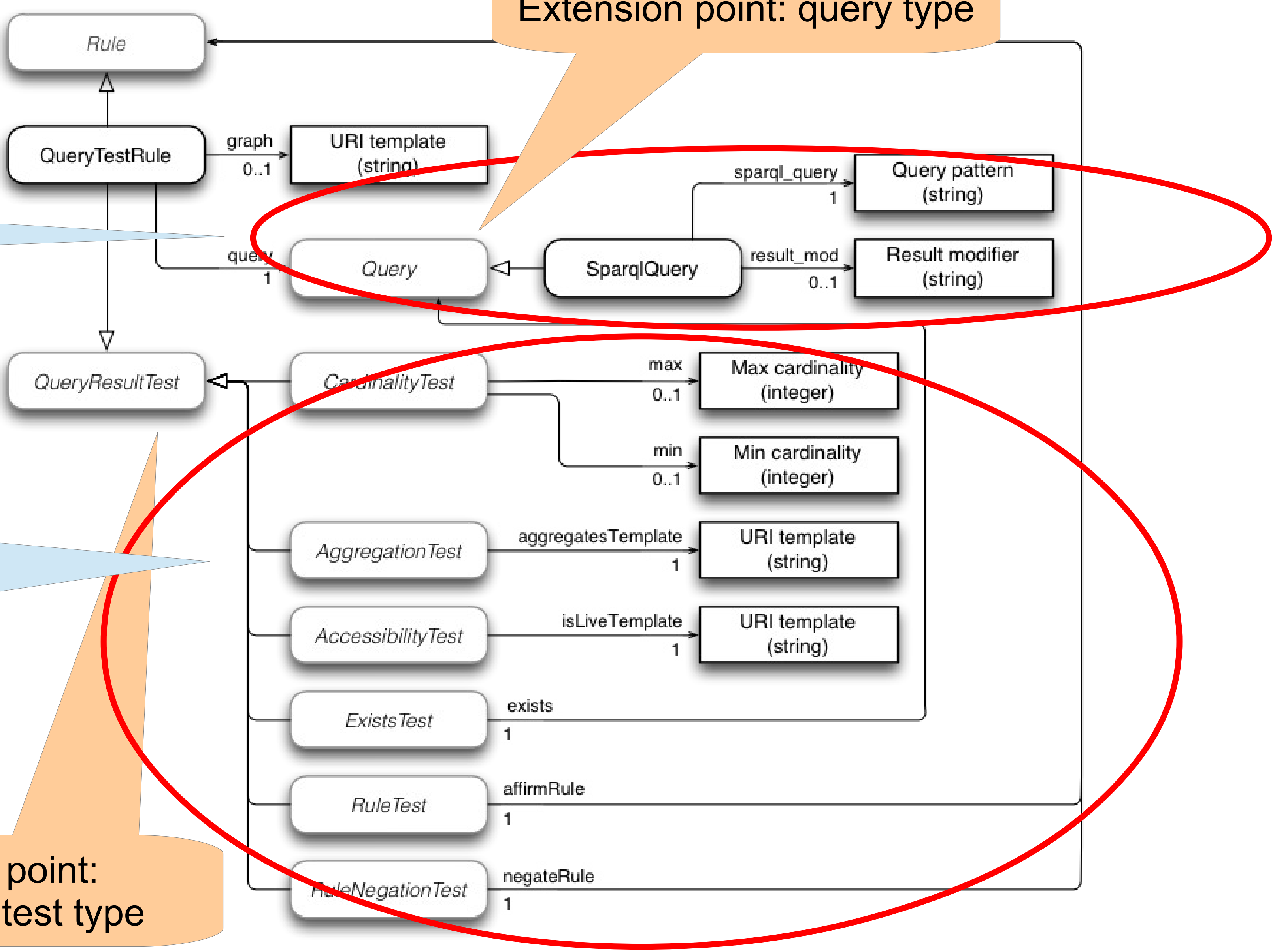
SHOULD contain one or more numeric ChemSpider identifiers

MAY include any number of chemical synonyms

The “Minim” model: checklists as linked data



The “Minim” model: query test rule as linked data



Extension point: query type

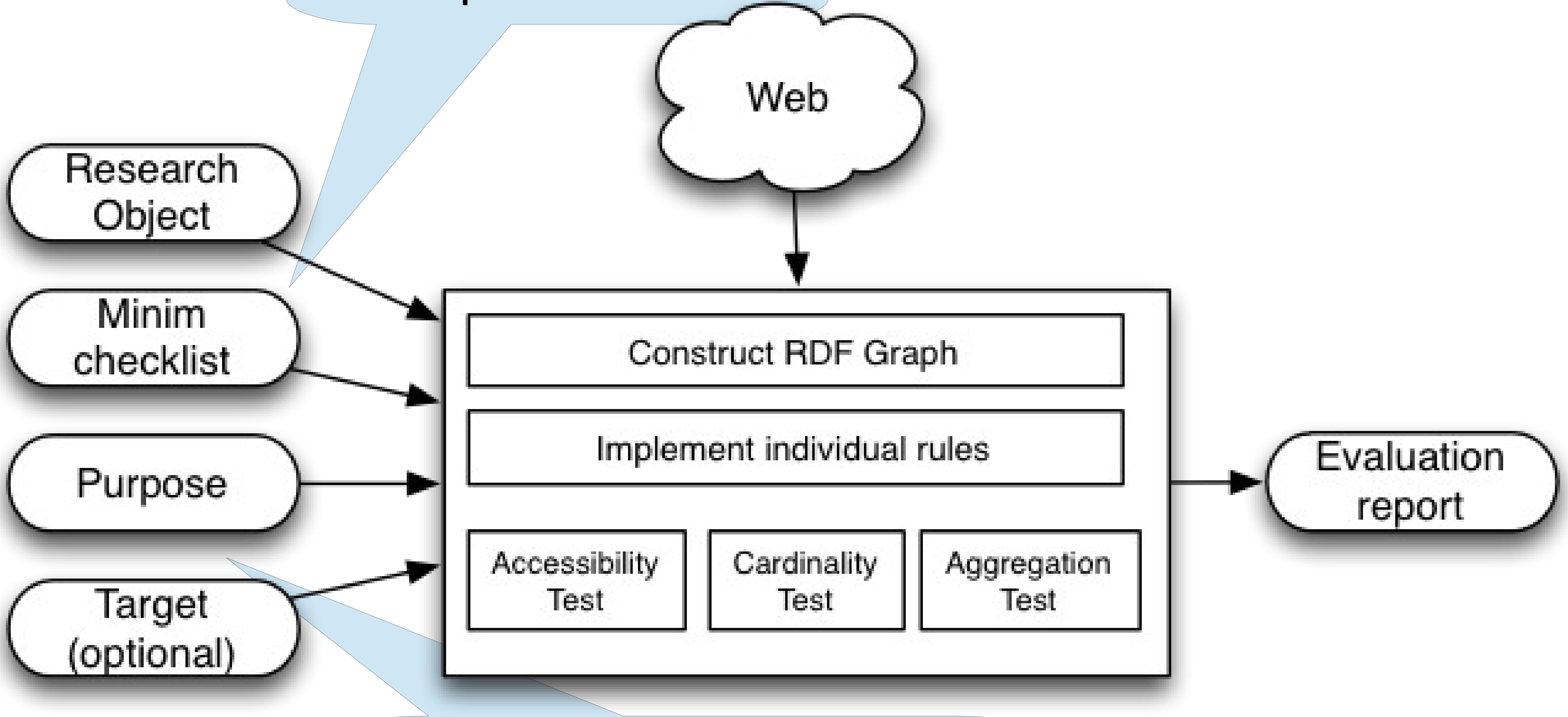
Query

Test query result

Extension point:
query result test type

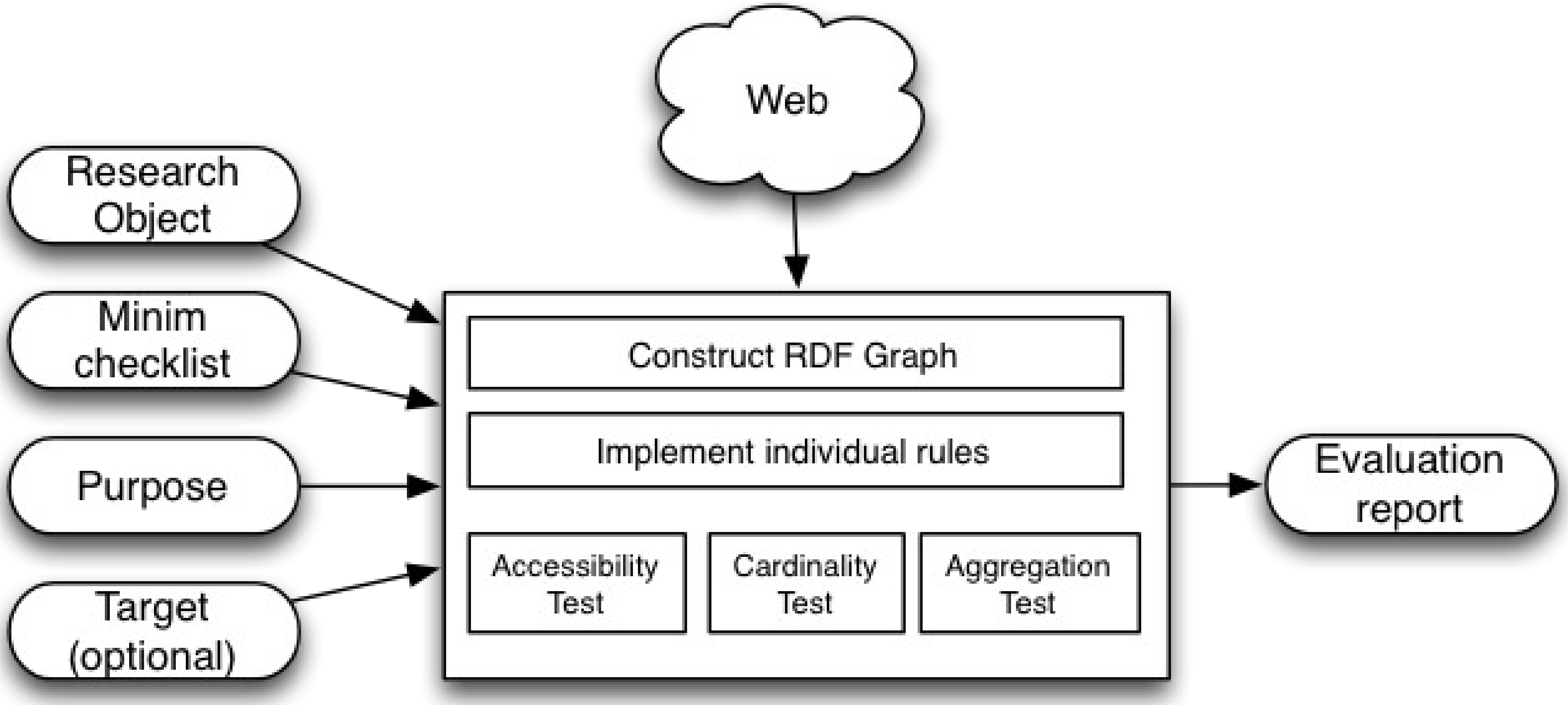
Evaluation overview (1) checklist selection

Minim file may contain multiple checklists



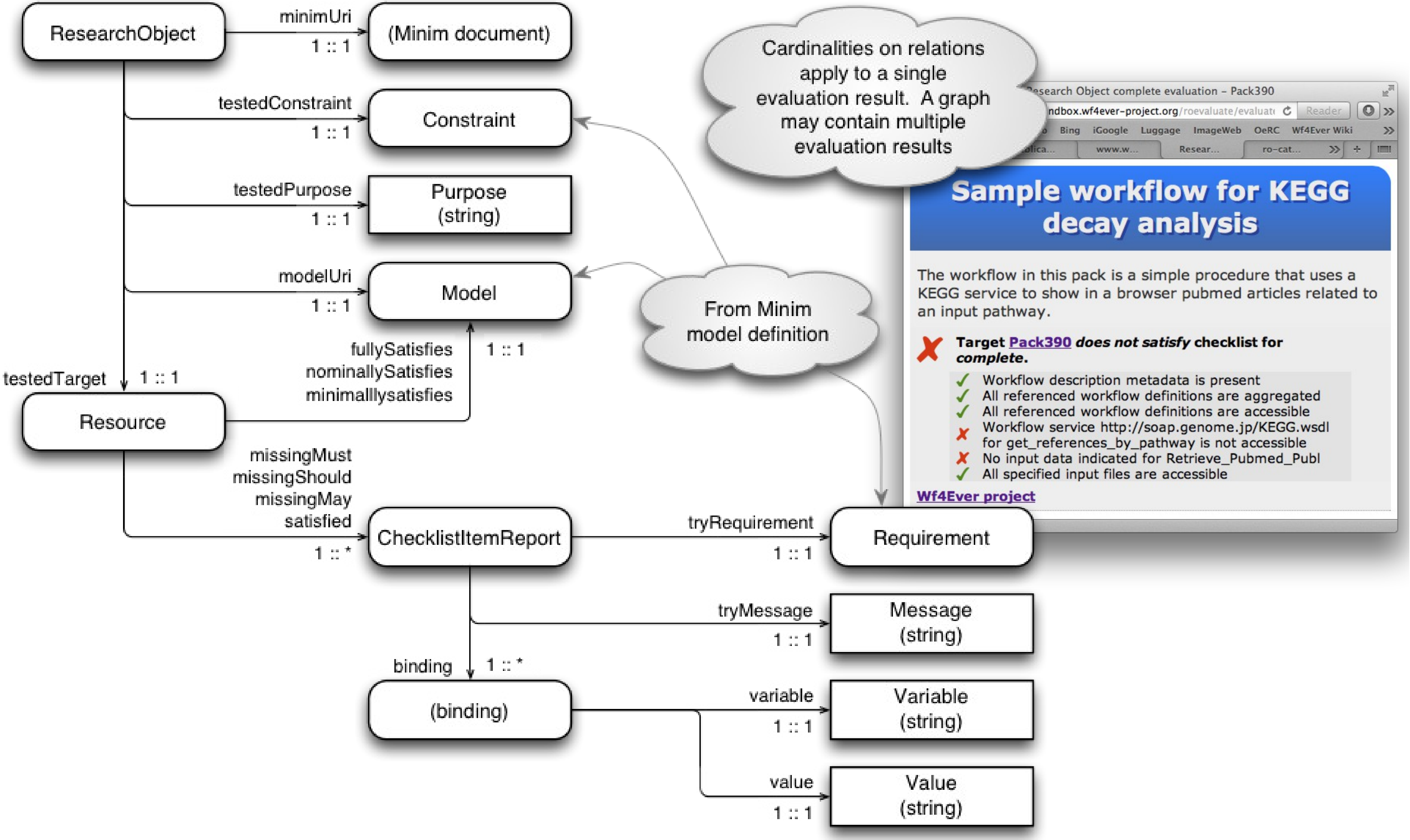
“Purpose” and “Target” used to select checklist to evaluate

Evaluation overview (2): checklist model evaluation



1. Construct RDF graph of RO annotations
2. Evaluate each requirement in the checklist model
3. Assemble result

Checklist evaluation: results as linked data



Cardinalities on relations apply to a single evaluation result. A graph may contain multiple evaluation results

From Minim model definition

Sample workflow for KEGG decay analysis

The workflow in this pack is a simple procedure that uses a KEGG service to show in a browser pubmed articles related to an input pathway.

✗ **Target Pack390 *does not satisfy* checklist for *complete*.**

- ✓ Workflow description metadata is present
- ✓ All referenced workflow definitions are aggregated
- ✓ All referenced workflow definitions are accessible
- ✗ Workflow service http://soap.genome.jp/KEGG.wsdl for get_references_by_pathway is not accessible
- ✗ No input data indicated for Retrieve_Pubmed_Publ
- ✓ All specified input files are accessible

Wf4Ever project

Evaluation and Reflections

Evaluation of approach

Our evaluation to date has focused on the capability of our model rather than its performance

- could checklists handle our Wf4Ever project requirements?
- how did our capabilities match those of other tools?

We report on:

- detection of workflow decay
- completeness of linked data chemical descriptions

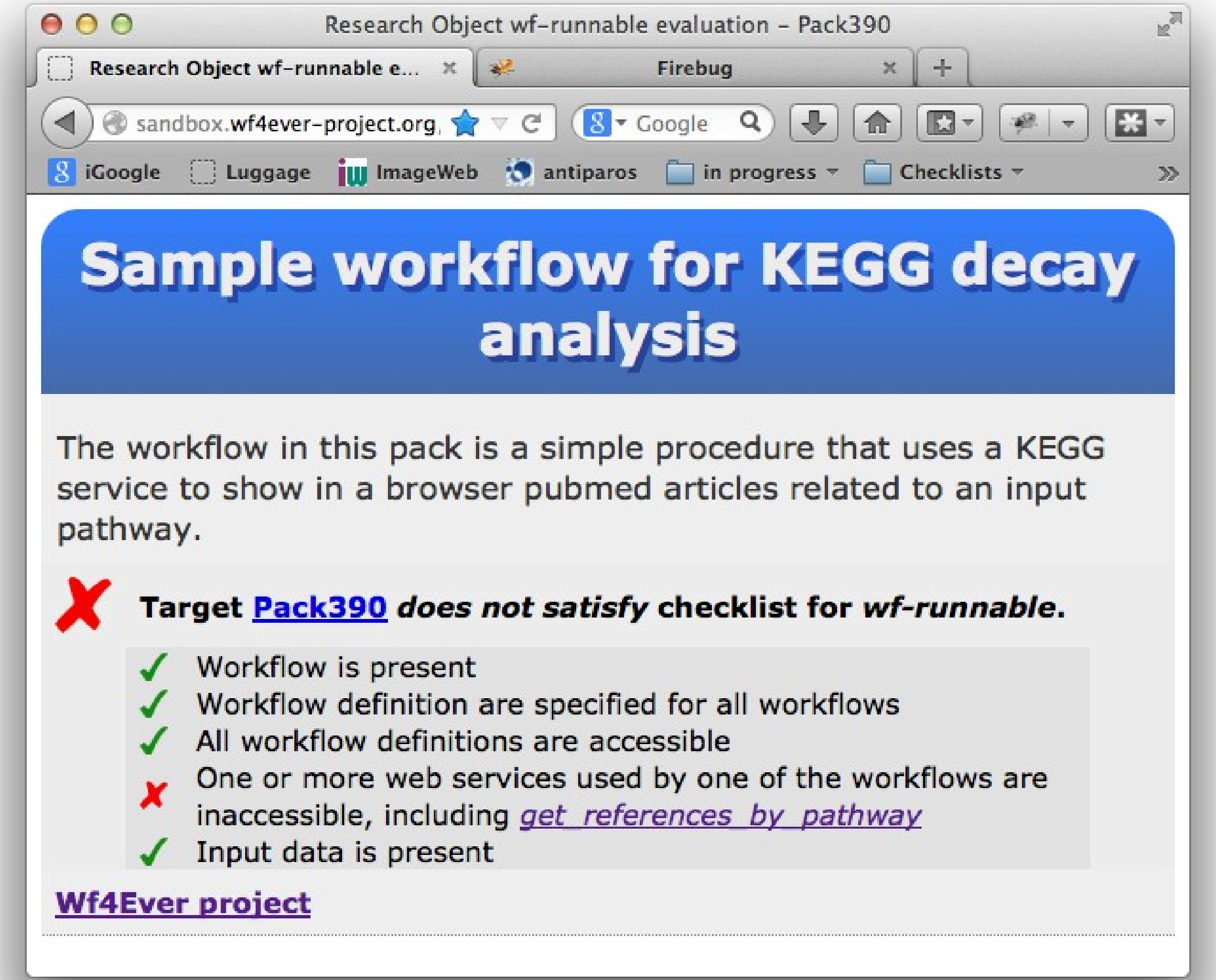
Workflow Decay Detection

2012: replacement of KEGG Web Services with REST services

Workflows located in myExperiment

Before shutdown, workflow runnability was confirmed

After shutdown, the checklist reports workflow decay


Research Object wf-runnable evaluation - Pack390
Research Object wf-runnable e...
Firebug
sandbox.wf4ever-project.org
Google
iGoogle
Luggage
ImageWeb
antiparos
in progress
Checklists

Sample workflow for KEGG decay analysis

The workflow in this pack is a simple procedure that uses a KEGG service to show in a browser pubmed articles related to an input pathway.

✗ Target Pack390 *does not satisfy* checklist for *wf-runnable*.

- ✓ Workflow is present
- ✓ Workflow definition are specified for all workflows
- ✓ All workflow definitions are accessible
- ✗ One or more web services used by one of the workflows are inaccessible, including *get_references_by_pathway*
- ✓ Input data is present

Wf4Ever project

http://sandbox.wf4ever-project.org/roevaluate/evaluate/trafficlight_html?minim=http://sandbox.wf4ever-project.org/rodl/ROs/Kegg-workflow-evaluation/Runnable-workflow-checklist.rdf&purpose=wf-runnable&RO=http://sandbox.wf4ever-project.org/rodl/ROs/Pack390/

Completeness of Chemical Descriptions

Requirements testing with SPIN

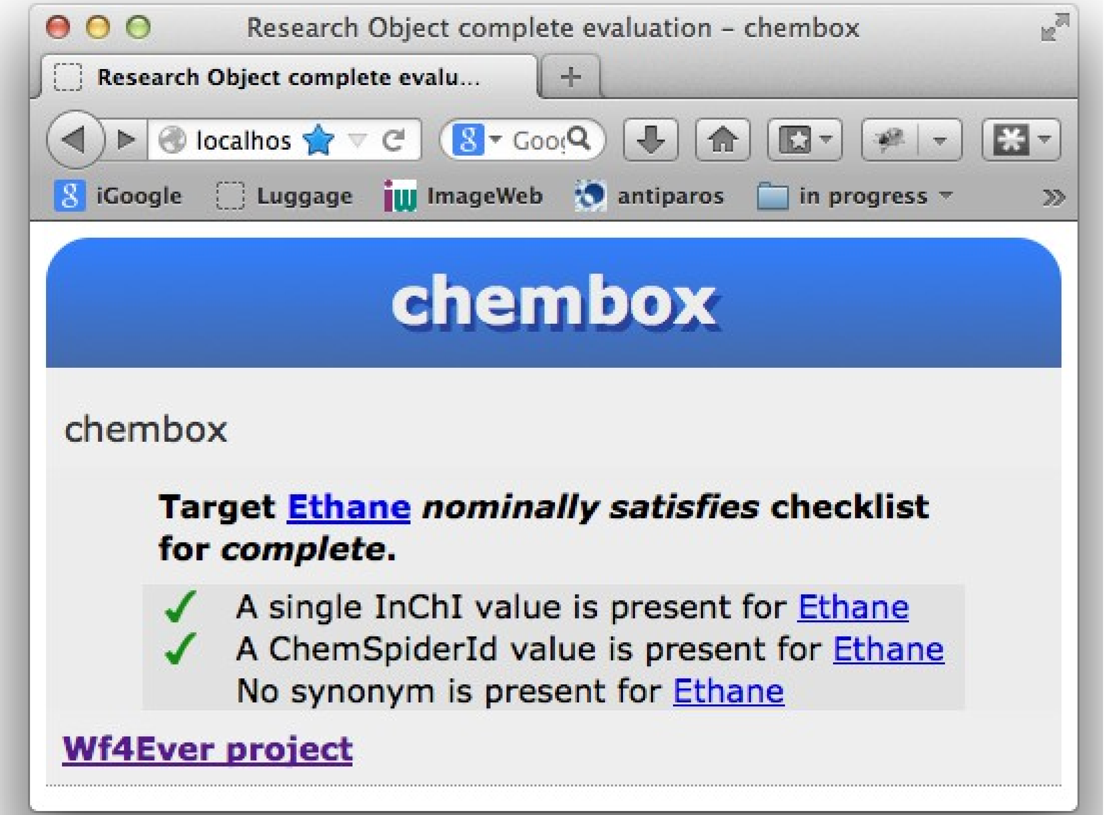
- Previous work, using SPARQL Inference Notation

“ChemBox” chemical descriptions extracted from Wikipedia

Key difference is query syntax:

- SPARQL vs SPIN

One test was unsuited to SPARQL query


Research Object complete evaluation – chembox
Research Object complete evalu...
localhos
Goo
iGoogle
Luggage
ImageWeb
antiparos
in progress

chembox

chembox

Target Ethane *nominally satisfies* checklist for *complete*.

✓ A single InChI value is present for Ethane
✓ A ChemSpiderId value is present for Ethane
No synonym is present for Ethane

Wf4Ever project

Comparison with OWL-based approach

(OWL – Web Ontology Language)

SPARQL lacks in-built inference capabilities

Some tests don't really work with open world assumption

Some tests look at more than just the data

- e.g. accessibility of web resource

OWL could be used in conjunction with the Minim model

SKOS Thesaurus quality

Finding Quality Issues in SKOS Vocabularies

- [Mader, et al]

Shows some gaps in our current checklist tool implementation

Gaps addressable using Minim model extensions

What can be checked?

- Obviously, not everything can be automated
- Starting from user requirements, we:
  - determine can be mechanized (possibly with additional annotations or provenance)
  - discuss how remaining evaluation can be performed (e.g. special application, manual review, etc.) and create annotations to indicate the outcomes
- This process may yield candidates for extending the checklist model

Granularity

- What is the granularity of checklist items?
  - whatever can be probed with a SPARQL query, down to the level of individual RDF triples.
- Some user requirements don't conveniently map down to this level
  - future work may consider model extensions for composing tests within a single checklist item
  - so far, this has not been a pressing requirement in our work, but the design is easy to imagine (e.g. logical combinations of individual tests).

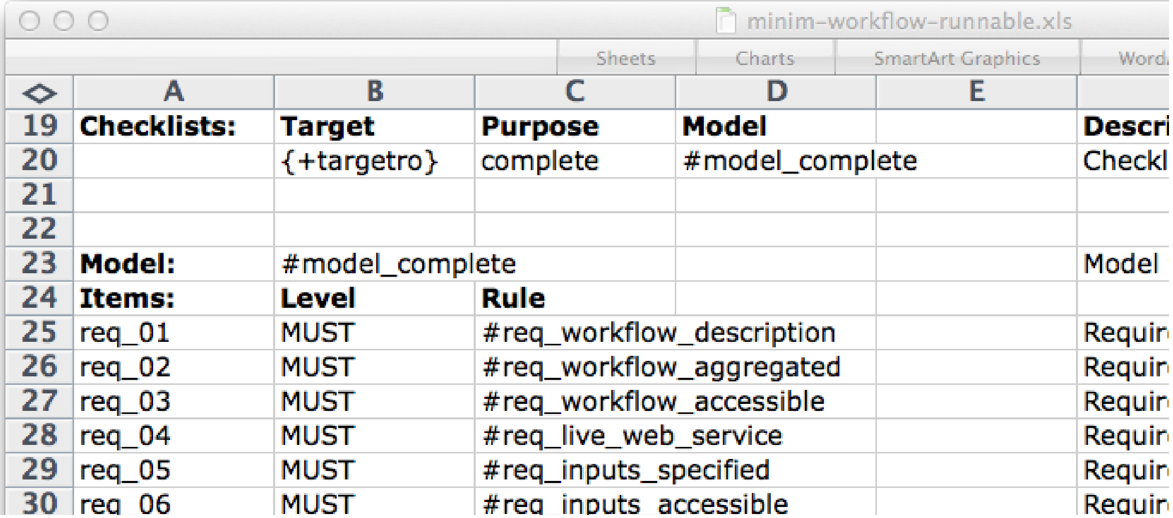
Performance and scalability

- Not yet formally evaluated
- But some Research Objects have proven slow to evaluate
  - Appears to be dominated by RDF load time
  - Performance problems have been overcome by using a lightweight RO creation service

Continuing Work
and
Concluding Remarks

Recent and ongoing work

Minim creation from spreadsheet



◇	A	B	C	D	E	
19	**Checklists:**	**Target**	**Purpose**	**Model**		**Descri**
20		{+targetro}	complete	#model_complete		Checkl
21						
22						
23	**Model:**	#model_complete				Model
24	**Items:**	**Level**	**Rule**			
25	req_01	MUST	#req_workflow_description			Requir
26	req_02	MUST	#req_workflow_aggregated			Requir
27	req_03	MUST	#req_workflow_accessible			Requir
28	req_04	MUST	#req_live_web_service			Requir
29	req_05	MUST	#req_inputs_specified			Requir
30	req_06	MUST	#req_inputs_accessible			Requir

Evaluating arbitrary linked data

- “Overlay RO” service
- lightweight ROs for linked data

Matching/aligning quality metrics with checklist capabilities

Checklist catalogue

Concluding remarks

Our goal: to assess the quality of information (data and computational methods) used and generated by researchers

We adopt checklists, which are a common tool for quality and safety assurance

Checklists are a pragmatic approach to assessing fitness-for-use, complementary to analysis of data quality dimensions

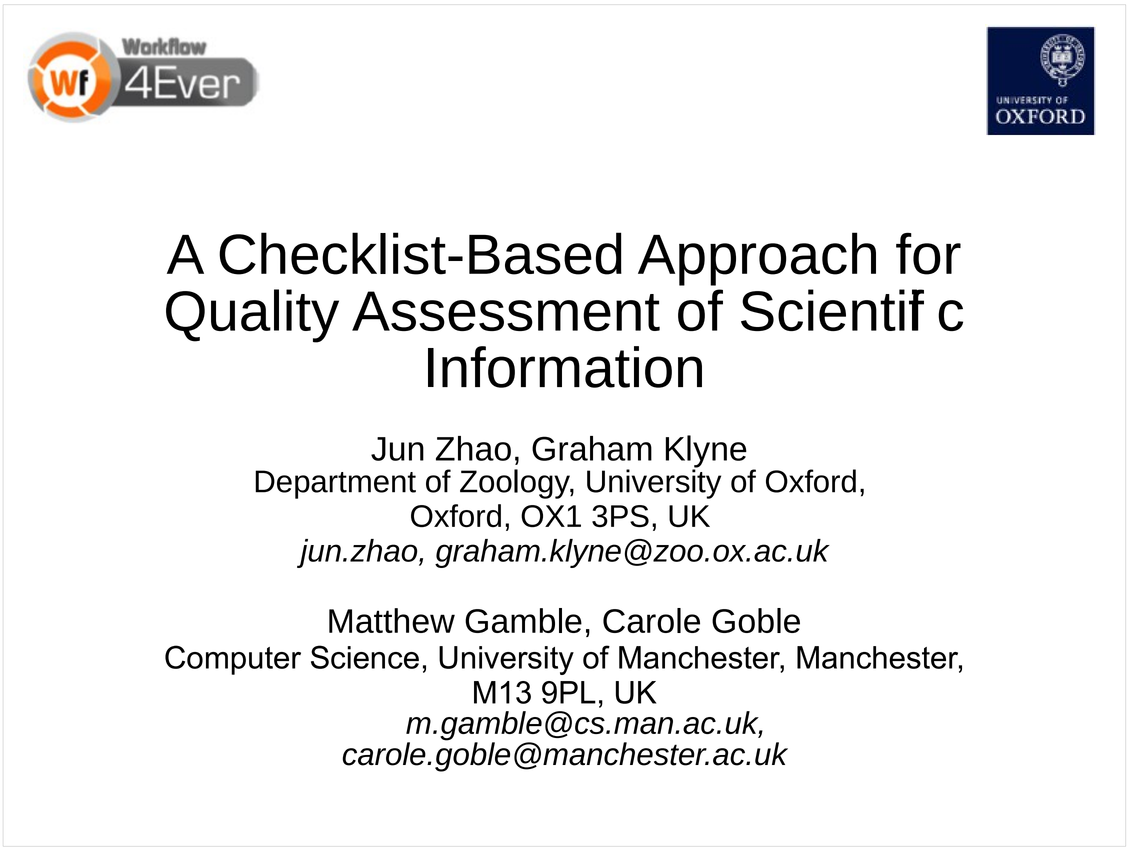
Our model allows automated tests to be combined with manual review

Acknowledgements

EU Wf4Ever project (270129)
funded under EU FP7 (ICT- 2009.4.1).
(http://www.wf4ever-project.org)

Links

- Paper
  - …
- Presentation
  - …
- Software
  - https://github.com/wf4ever/ro-manager
- Evaluation scripts and data
  - https://github.com/wf4ever/ro-catalogue/tree/master/v0.1/minim-evaluation

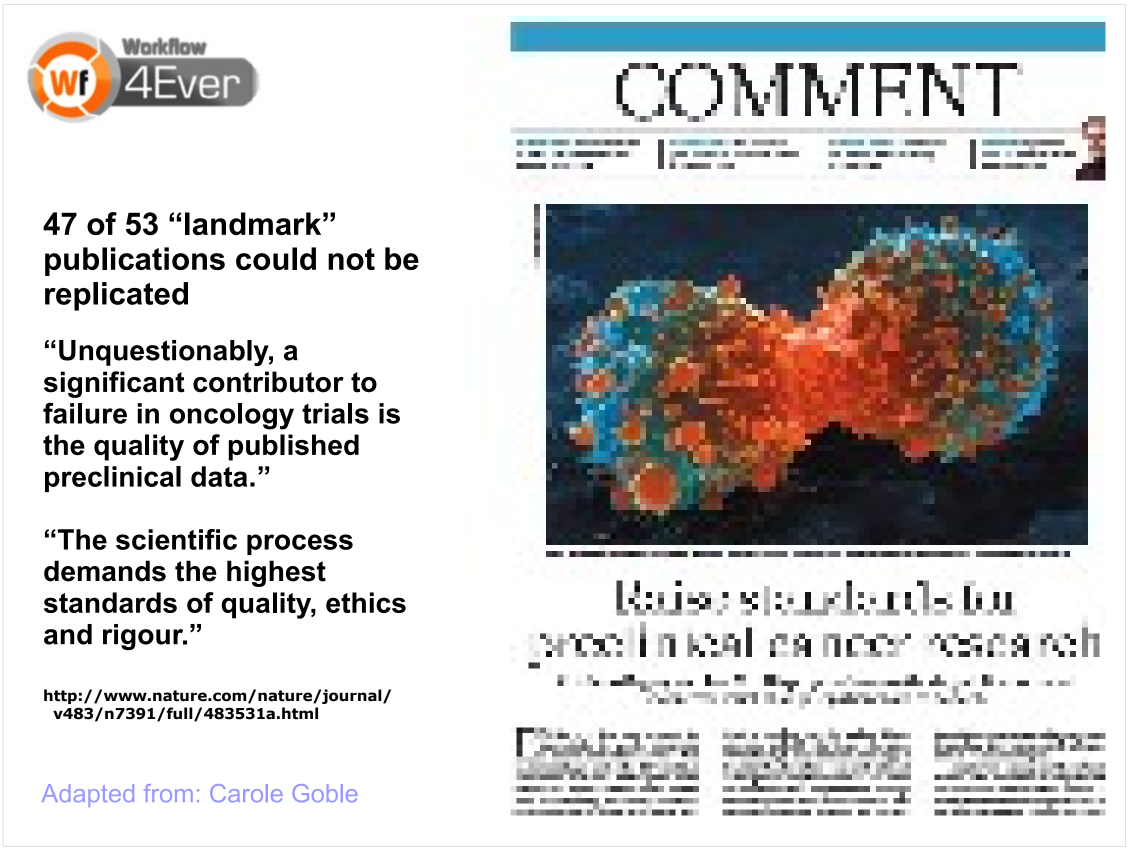


A Checklist-Based Approach for Quality Assessment of Scientific Information

Jun Zhao, Graham Klyne
Department of Zoology, University of Oxford,
Oxford, OX1 3PS, UK
jun.zhao, graham.klyne@zoo.ox.ac.uk

Matthew Gamble, Carole Goble
Computer Science, University of Manchester, Manchester,
M13 9PL, UK
m.gamble@cs.man.ac.uk,
carole.goble@manchester.ac.uk

I shall talk today about our checklist based approach to scientific information quality assessment...



Workflow 4Ever

47 of 53 “landmark” publications could not be replicated

“Unquestionably, a significant contributor to failure in oncology trials is the quality of published preclinical data.”

“The scientific process demands the highest standards of quality, ethics and rigour.”

http://www.nature.com/nature/journal/v483/n7391/full/483531a.html

Adapted from: Carole Goble

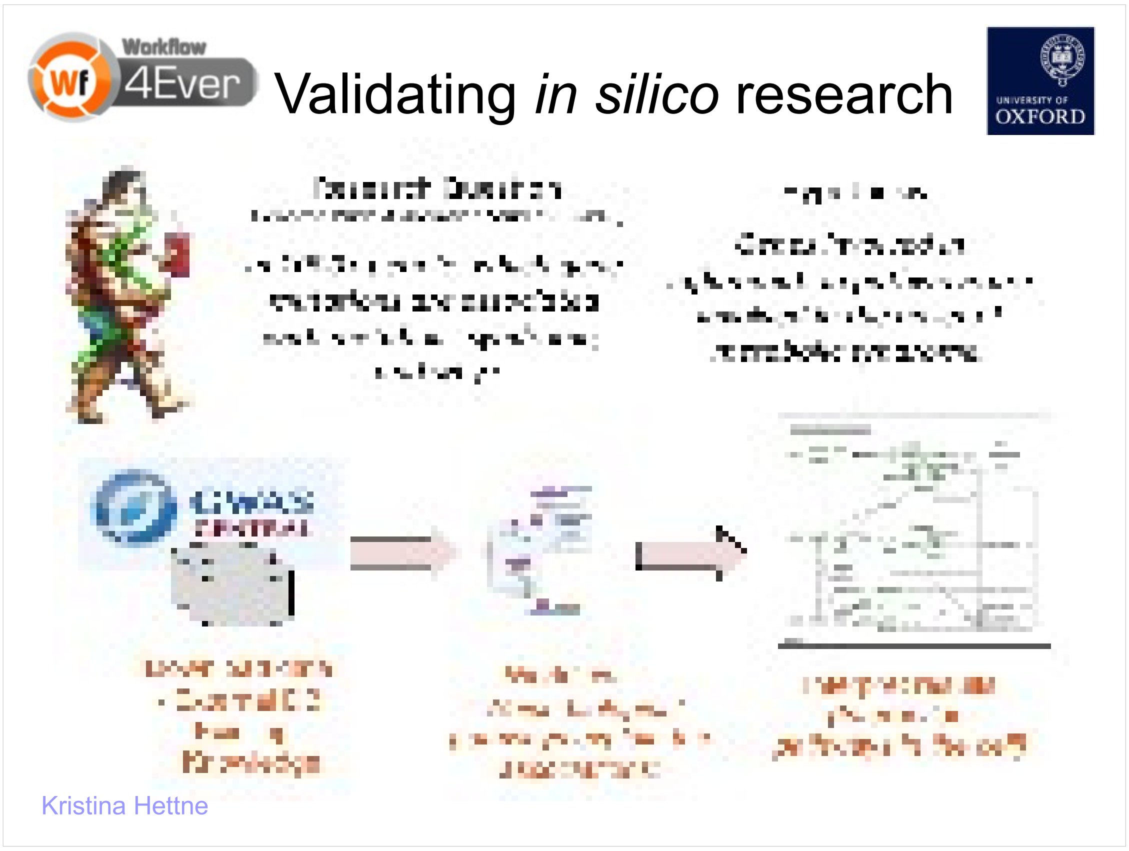
Science proceeds by building on the research of others, and the quality of published work is key to supporting continued progress.

But this study published in Nature found the results of 47 out of 53 “landmark” studies in preclinical cancer research could not be replicated, casting doubts on their suitability as a basis for further research.

“Landmark” here means results that are regarded as reference points by a research community, and which are widely used to underpin ongoing work.

This reflects a concern among scientists about quality of published research [1], and the high costs of basing further work on poor results.

[1] e.g., see http://retractionwatch.wordpress.com



Validating *in silico* research

Kristina Hettne

Science is increasingly dependent on computed results and *in silico* investigations, such as this genome wide association study into mechanisms involved in metabolic syndrome. Our Wf4Ever project is concerned with conservation of scientific workflows used, and with mitigation of workflow decay.

Such research is based on processes that take place in the hidden recesses of computer systems, so how are we to judge the reliability of the results unless we can assess the quality of the input data and of the workflows used?

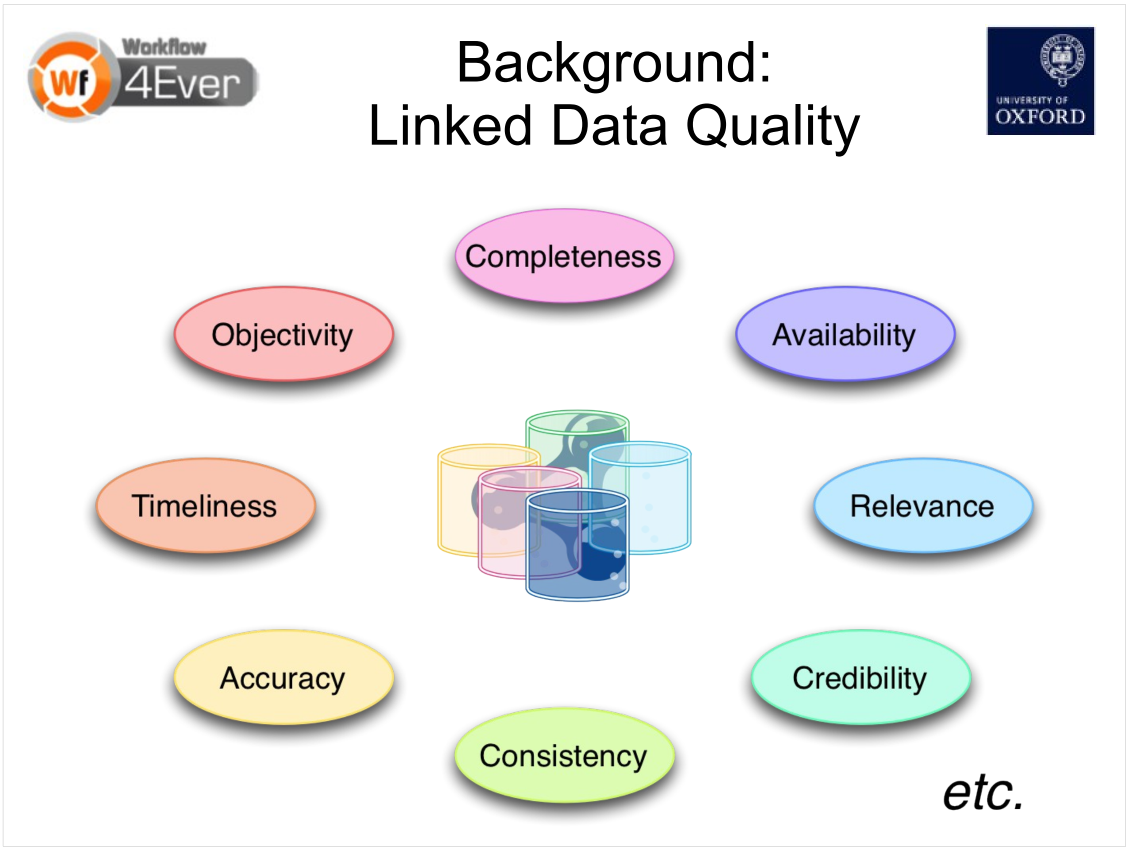
The concern for quality thus extends to the underpinnings of *in silico* investigations [1]. Our work aims to support researchers in the creation of high quality *in silico* research outputs, and in the selection of resources upon which they base their work.

[1] ttp://www.nature.com/news/mozilla-plan-seeks-to-debug-scientific-code-1.13812

The Goal of Our Work

We aim to provide tools that assess
the ***quality of information***
(***data*** and ***computational artifacts***)
used and generated by researchers

As such we are concerned with assessing the **quality of information** – both data, and computational artifacts – used and generated by informatics-based researchers.



Background: Linked Data Quality

Completeness

Objectivity

Availability

Timeliness

Relevance

Accuracy

Credibility

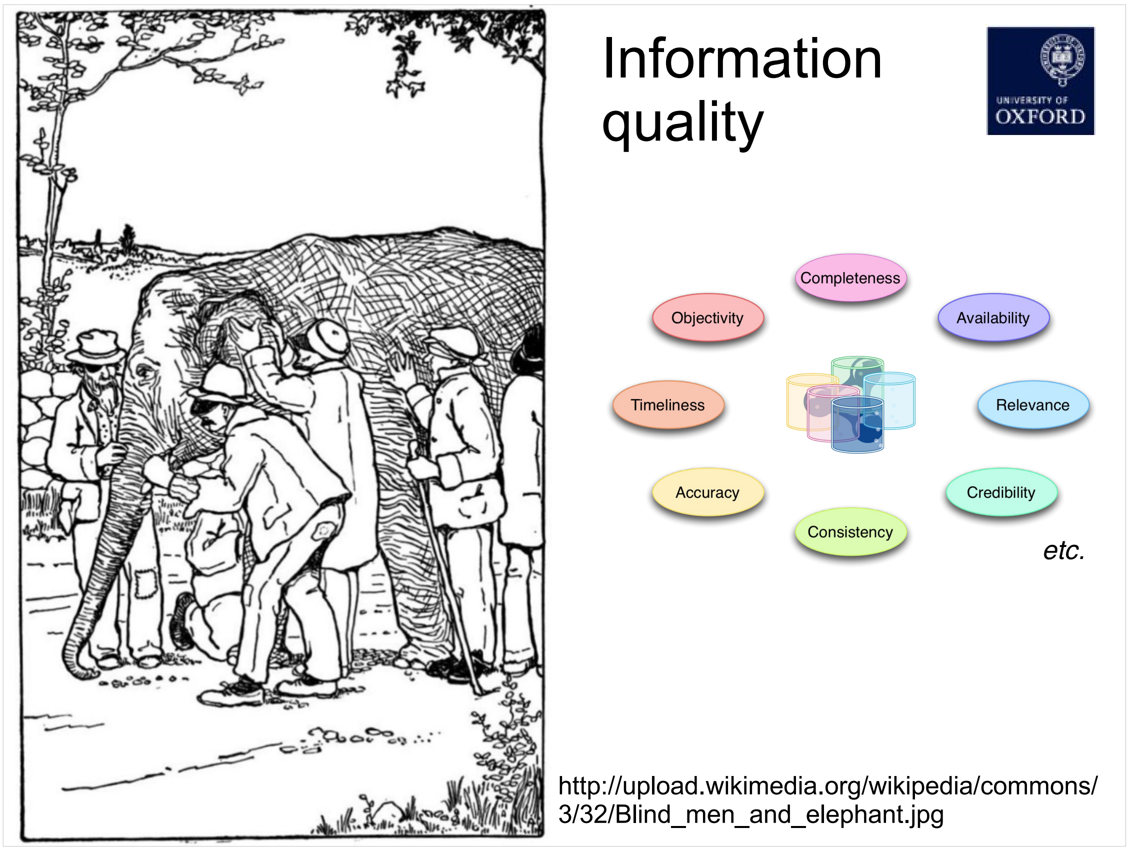
Consistency

etc.

Information quality assessment has been extensively studied in management science, and web-based information systems, to assure the quality of manufactured and information products.

Most existing approaches to quality assessment use quality metrics, and produce an overall quality measure by integrating over a number of quality dimensions, such as accuracy, completeness, or credibility.

((These approaches are seen in work on linked data quality from a number of researchers, such as Bizer, Zaveri and others.))



Information quality

Completeness

Objectivity

Availability

Timeliness

Relevance

Accuracy

Credibility

Consistency

etc.

http://upload.wikimedia.org/wikipedia/commons/3/32/Blind_men_and_elephant.jpg

But this focus on the dimensions of information quality can leave one feeling like the blind men of legend, trying to evaluate an elephant by touch; its leg like a pillar, its ear like a fan, its tusk like a solid pipe, etc.

These are all features of an elephant that may be assessed and described, but it can be difficult to see how they relate to the elephant as a whole, and in particular to how well it may perform in any given circumstance ...

For these elephant *users*...

http://www.flickr.com/photos/leosoueu/7121406603/

But for these passengers, these *users* of an elephant, I would expect their main concern here would be...

For these elephant *users*...

… “can this elephant get me across this river?”

http://www.flickr.com/photos/leosoueu/7121406603/

… “can this elephant get us across this river?”

Our approach to information quality assessment

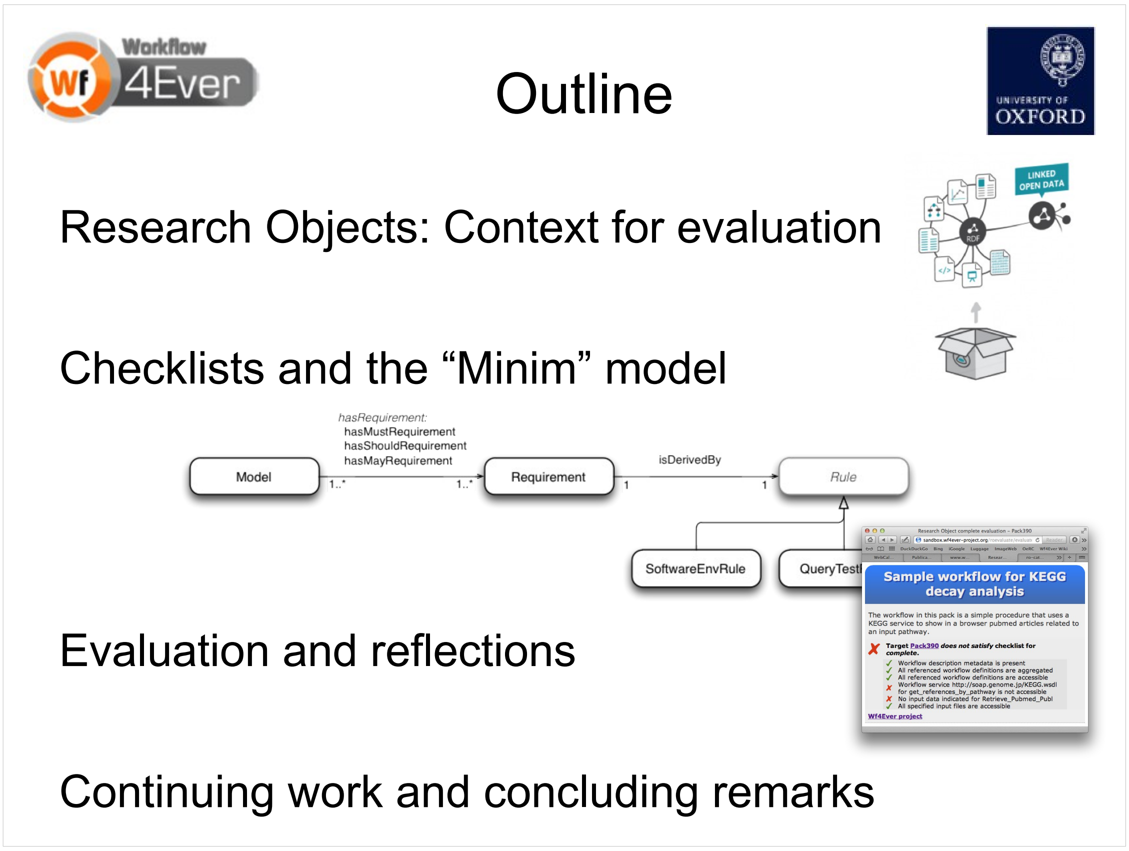
We see complementary approaches to quality:

- An analysis of what constitutes quality, and
- Users' concern about fitness for use

The work presented here is focused on the latter of these concerns

I offer this as an illustration of two complementary approaches to information quality assessment:
 - an analysis of what constitutes quality, and
 - users' concern about fitness for use

Our work reported here focuses on the fitness-for-use aspect of quality assessment.



Outline

Research Objects: Context for evaluation

Checklists and the “Minim” model

Sample workflow for KEGG decay analysis

The workflow in this pack is a simple procedure that uses a KEGG service to show in a browser pubmed articles related to an input pathway.

✗ **Target Pack390 *does not satisfy* checklist for *complete*.**

- ✓ Workflow description metadata is present
- ✓ All referenced workflow definitions are aggregated
- ✓ All referenced workflow definitions are accessible
- ✗ Workflow service http://soap.genome.jp/KEGG.wsdl for get_references_by_pathway is not accessible
- ✗ No input data indicated for Retrieve_Pubmed_Publ
- ✓ All specified input files are accessible

Wf4Ever project

Evaluation and reflections

Continuing work and concluding remarks

The rest of this presentation will address the following topics:

- Our use of Research Objects to define the context, or scope, of a quality evaluation

- Then, I'll introduce checklists, our “Minim” model for describing them as linked data, and associated tools

- Next, I'll describe our evaluation of this approach, and address some comments and questions that have come up in the process

- And finally, I'll wrap up with mention of our continuing work, and some concluding remarks

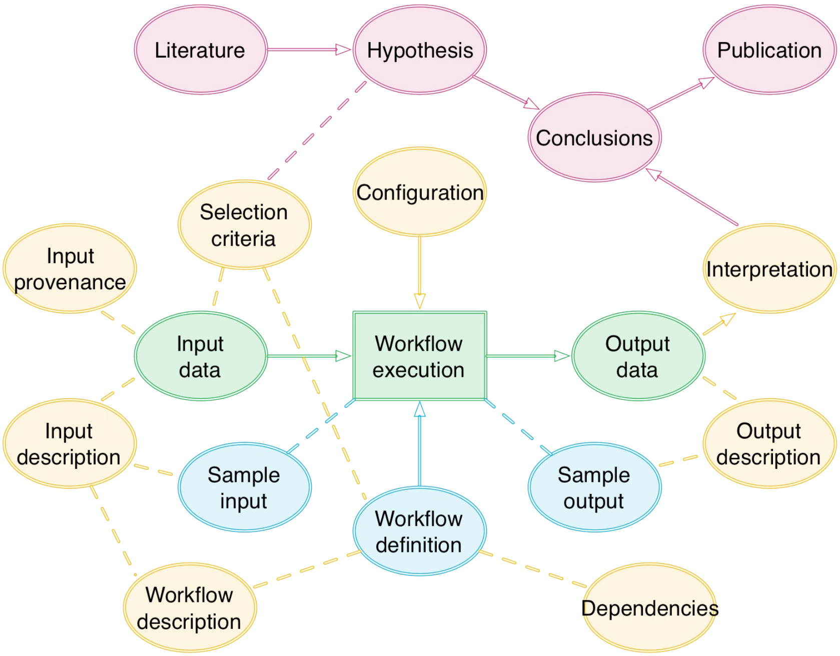
Workflow 4Ever

UNIVERSITY OF OXFORD

Research Objects

Context for information quality evaluation

Context for evaluation of scientific information quality

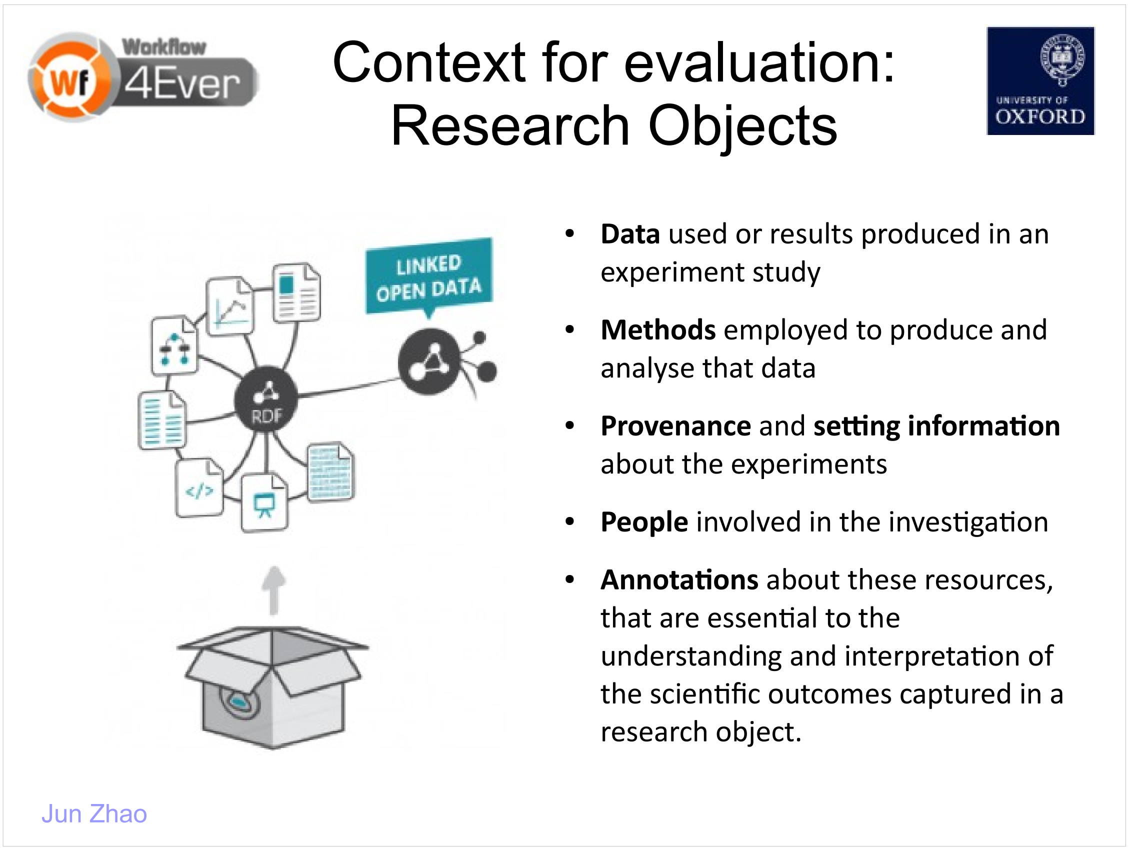


Many aspects of quality evaluation are performed with reference to some context or scope of use.

In general, we do not evaluate research artifacts in isolation, but as part of some investigation involving a constellation of related artifacts.

The suitability of any artifact may depend on its role with respect to the other artifacts that are also part of the investigation.

This collection of related artifacts constitutes the context of an investigation, or experiment, and it is within this context that we seek to apply our quality assessments.



Context for evaluation: Research Objects

- **Data** used or results produced in an experiment study
- **Methods** employed to produce and analyse that data
- **Provenance** and **setting information** about the experiments
- **People** involved in the investigation
- **Annotations** about these resources, that are essential to the understanding and interpretation of the scientific outcomes captured in a research object.

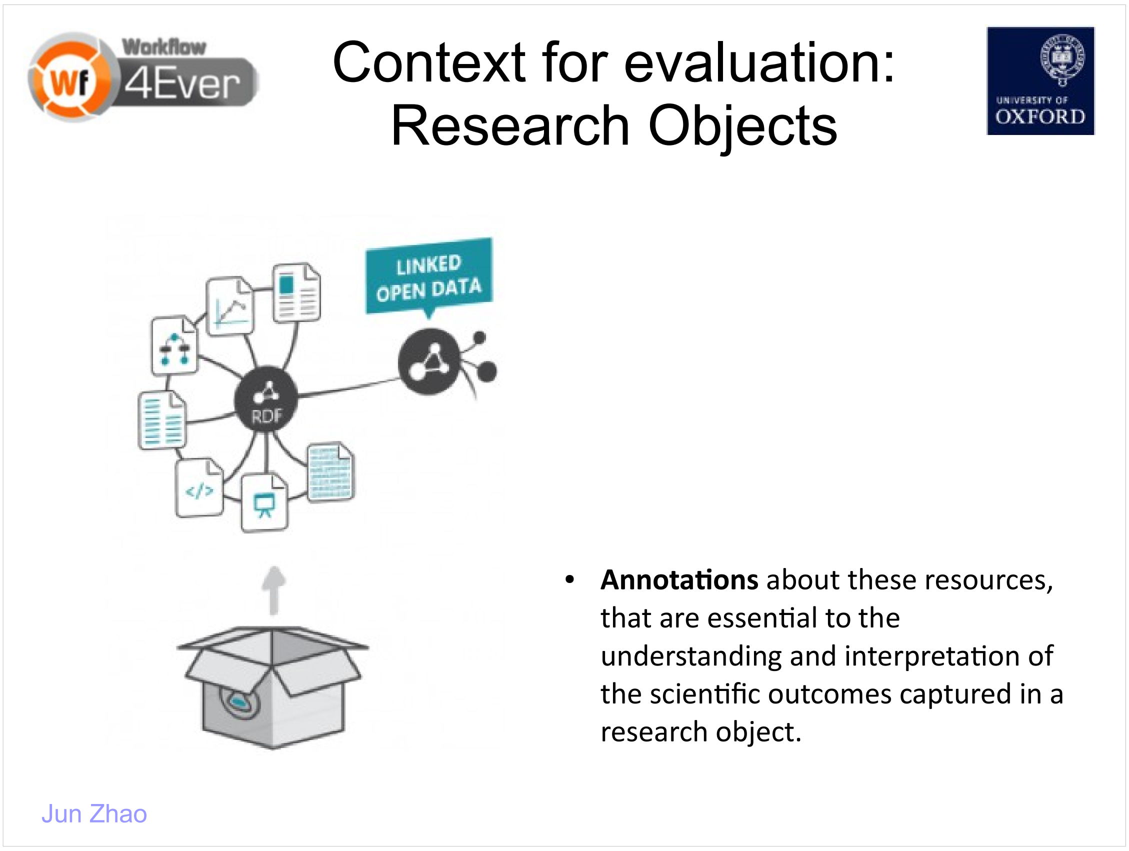
Jun Zhao

We use “Research Objects”, or ROs, to represent this context in the form of an aggregation of related resources, encapsulating essential information needed to understand, reproduce and re-use its elements.

The RO model is based on existing standards such as Object Re-use and Exchange (ORE), Annotation Ontology (AO), W3C Provenance (PROV), etc.

If we look inside an RO, we may find information about:

- data used
- results generated
- descriptions of methods
- provenance and configuration
- people involved
- *and...*



Context for evaluation: Research Objects

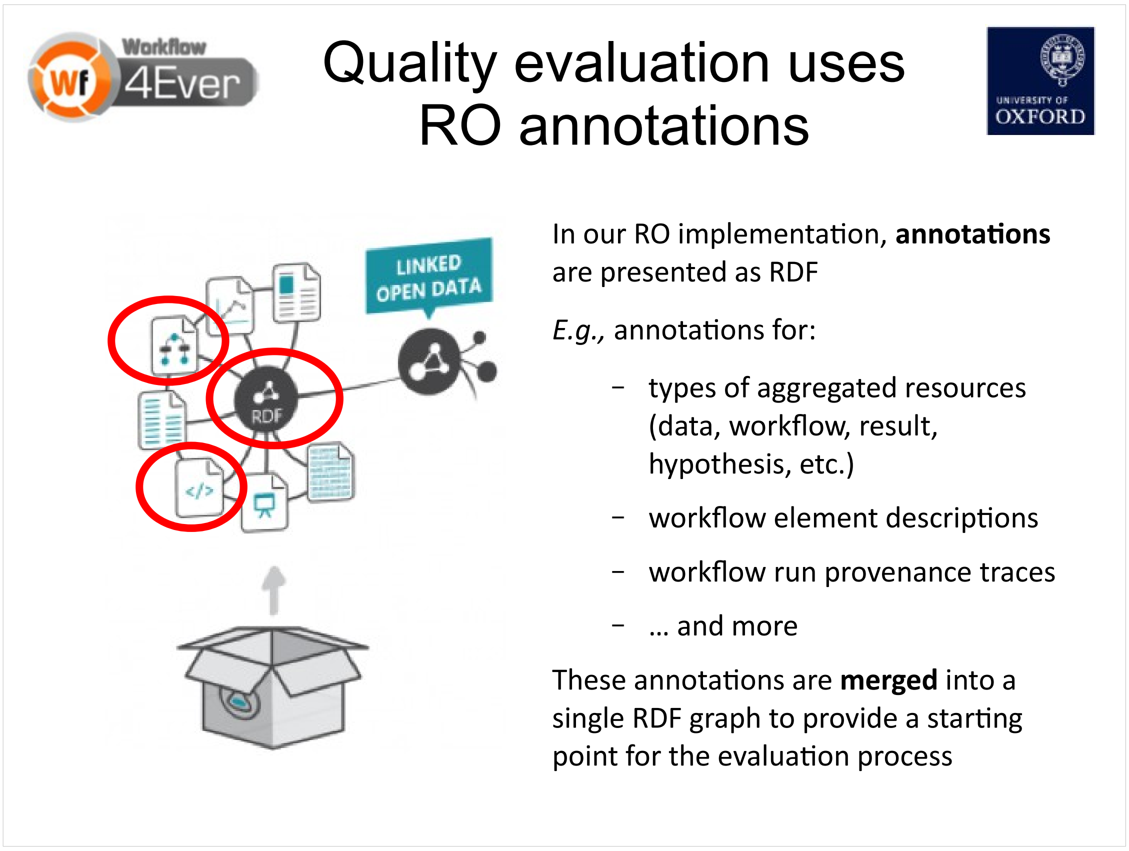
- **Annotations** about these resources, that are essential to the understanding and interpretation of the scientific outcomes captured in a research object.

Jun Zhao

… annotations about any of these

These annotations provide additional information needed to understand the artifacts and properly interpret their significance.

In our work, we use an RO as a container for information that will be used in assessing the quality of an artifact.



Quality evaluation uses RO annotations

In our RO implementation, **annotations** are presented as RDF

E.g., annotations for:

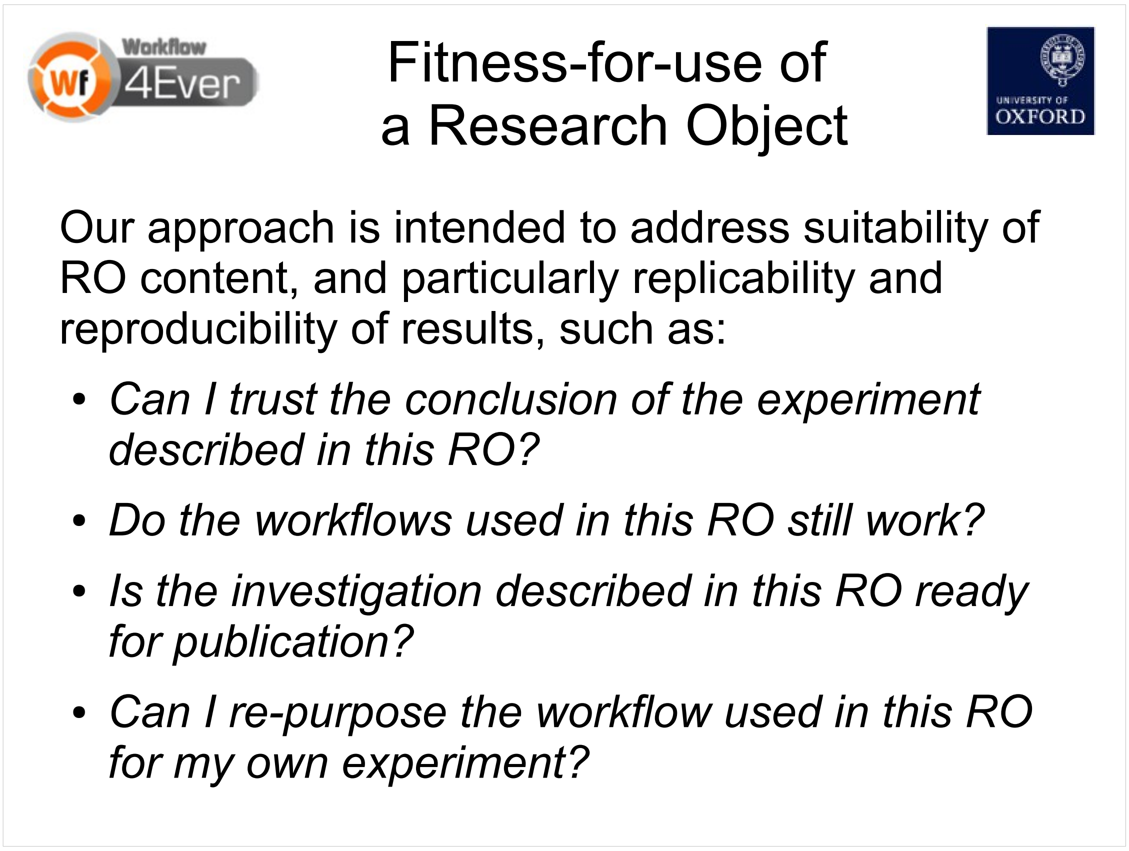
- types of aggregated resources (data, workflow, result, hypothesis, etc.)
- workflow element descriptions
- workflow run provenance traces
- … and more

These annotations are **merged** into a single RDF graph to provide a starting point for the evaluation process

Our *implementation* of Research Objects uses RDF to represent annotations, covering details such as....

- types of resources encapsulated in the RO
- descriptions of workflow components
- provenance traces of workflow runs

There are no constraints on the kind of information that can be provided, and our evaluation tools are designed to be able to work with whatever information is available.



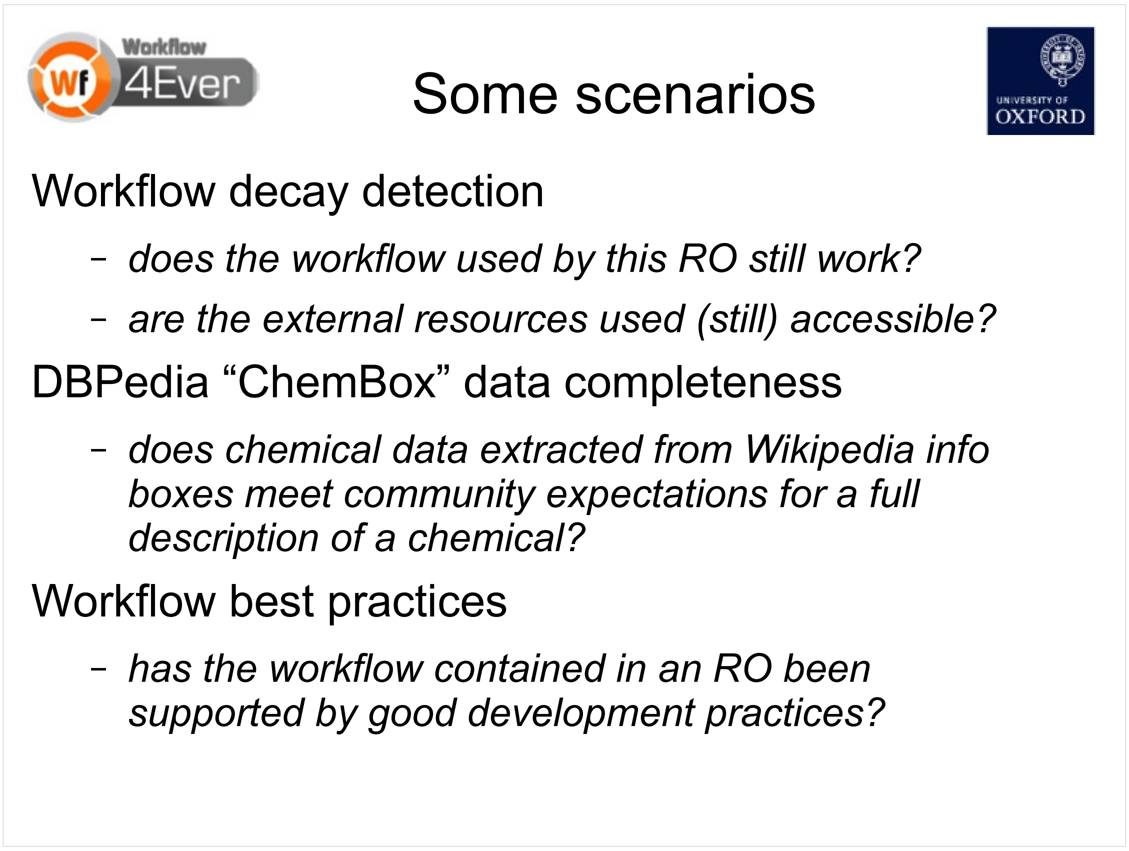
Fitness-for-use of a Research Object

Our approach is intended to address suitability of RO content, and particularly replicability and reproducibility of results, such as:

- *Can I trust the conclusion of the experiment described in this RO?*
- *Do the workflows used in this RO still work?*
- *Is the investigation described in this RO ready for publication?*
- *Can I re-purpose the workflow used in this RO for my own experiment?*

In determining the fitness-for-use of a Research Object, we aim to answer questions like:

- can I trust the conclusions presented?
- do the workflows used still work?
- is the investigation described ready for publication?
- can I re-use the workflow or other elements in my own work?
- *etc.*



Some scenarios

Workflow decay detection

- *does the workflow used by this RO still work?*
- *are the external resources used (still) accessible?*

DBPedia “ChemBox” data completeness

- *does chemical data extracted from Wikipedia info boxes meet community expectations for a full description of a chemical?*

Workflow best practices

- *has the workflow contained in an RO been supported by good development practices?*

Here are some particular scenarios we have been working with:

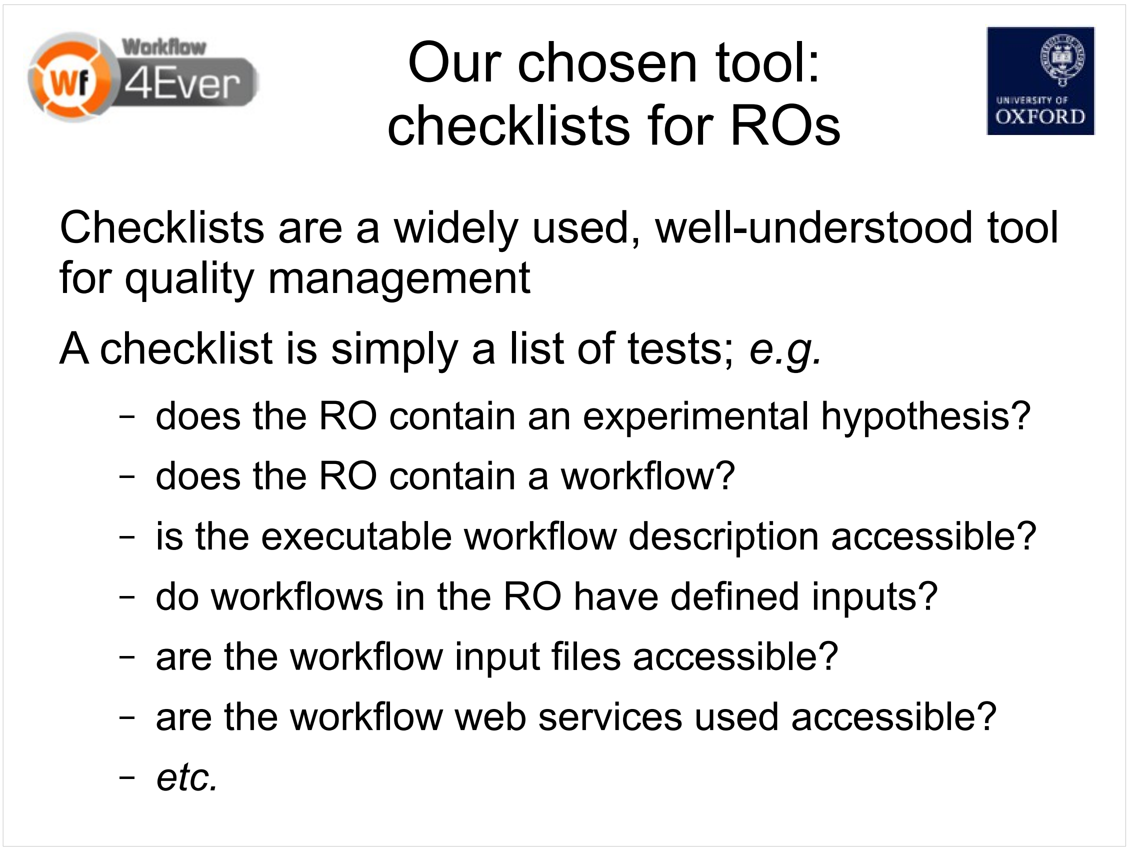
Workflow decay detection, e.g.:
 - does a (once-working) workflow still work; can its results be replicated by re-running the workflow
 - can it be run on different data to compare results?
 - are the external services used and other resources needed still available for use?

Another scenario deals with completeness of information... does chemical data extracted from Wikipedia meet the chemistry research community's expectations for a complete chemical description?

We have also looked at best practices... has a workflow described in an RO been developed in accord with community best practices?

Checklists and the “Minim” model

I shall now introduce our Minim model



Our chosen tool: checklists for ROs

Checklists are a widely used, well-understood tool for quality management

A checklist is simply a list of tests; *e.g.*

- does the RO contain an experimental hypothesis?
- does the RO contain a workflow?
- is the executable workflow description accessible?
- do workflows in the RO have defined inputs?
- are the workflow input files accessible?
- are the workflow web services used accessible?
- *etc.*

Our chosen tool for addressing quality questions raised is the humble checklist.

I expect everyone here has come across checklists in some form or another – they are widely used as tools for quality management and safety assurance.

In summary, a checklist is simply a list of requirements that we would wish to be satisfied.

Some specific requirements we have considered include:

- does an RO contain an experimental hypothesis?
- does it contain an abstract workflow description?
- is there an accessible, executable definition for all workflows used?
- *etc.*

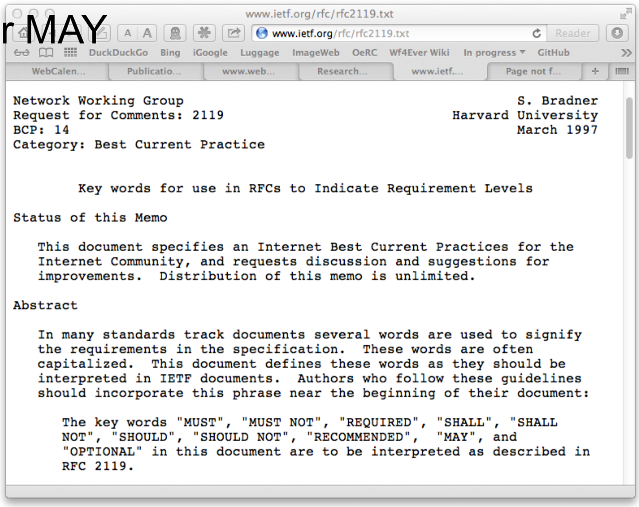
Requirement levels

Each checklist test has a requirement level:

- MUST, SHOULD or MAY
- from RFC 2119

The overall checklist result reflects the level of satisfied and unsatisfied requirements:

e.g. a failed MUST is more serious than a failed SHOULD



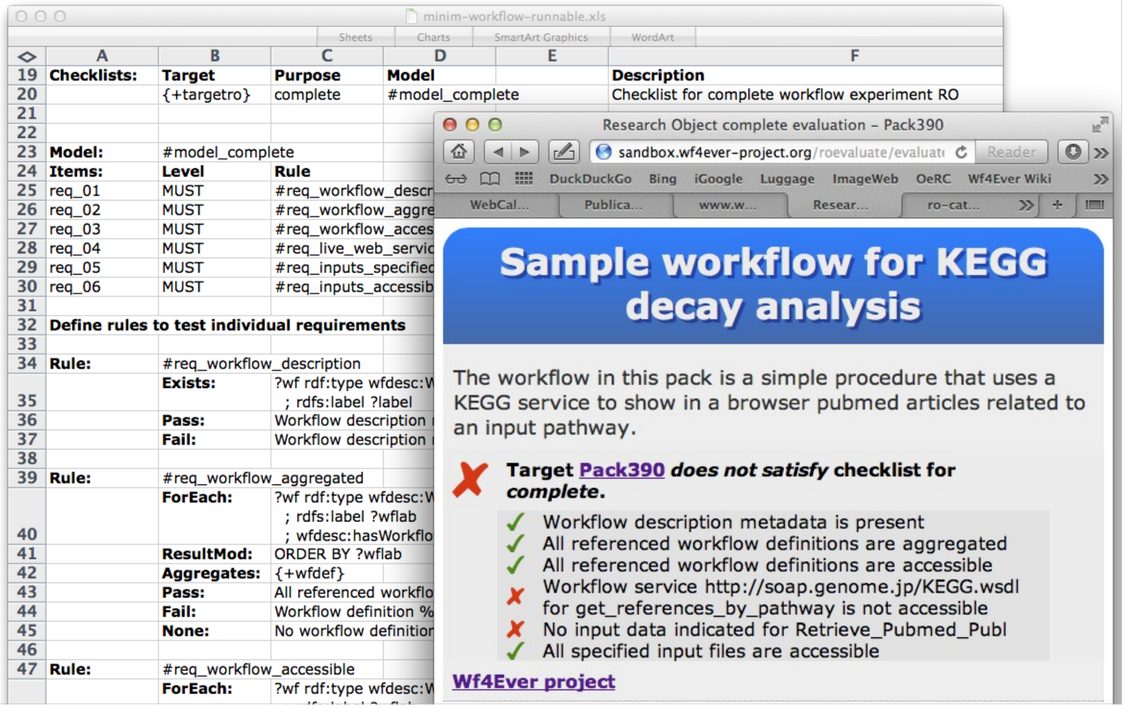
```
Network Working Group                                         S. Bradner
Request for Comments: 2119                            Harvard University
BCP: 14                                                       March 1997
Category: Best Current Practice


        Key words for use in RFCs to Indicate Requirement Levels

Status of this Memo

   This document specifies an Internet Best Current Practices for the
   Internet Community, and requests discussion and suggestions for
   improvements.  Distribution of this memo is unlimited.

Abstract

   In many standards track documents several words are used to signify
   the requirements in the specification.  These words are often
   capitalized.  This document defines these words as they should be
   interpreted in IETF documents.  Authors who follow these guidelines
   should incorporate this phrase near the beginning of their document:

      The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL
      NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED",  "MAY", and
      "OPTIONAL" in this document are to be interpreted as described in
      RFC 2119.
```

Associated with each checklist requirement is a requirement level: MUST, SHOULD or MAY. This idea is borrowed from IETF practice for defining Internet technical standards.

The overall result of evaluating a checklist reflects the levels of individual satisfied and unsatisfied requirements. For example, a failing MUST requirement is a more serious problem than a failing SHOULD or MAY requirement.

(The overall result is reflected as:
- fully satisfied
- nominally satisfied
- minimally satisfied
- not satisfied)

Runnable workflow checklist



	A	B	C	D	E	F
19	**Checklists:**	**Target**	**Purpose**	**Model**		**Description**
20		{+targetro}	complete	#model_complete		Checklist for complete workflow experiment RO
21						
22						
23	**Model:**	#model_complete				
24	**Items:**	**Level**	**Rule**			
25	req_01	MUST	#req_workflow_descr			
26	req_02	MUST	#req_workflow_aggre			
27	req_03	MUST	#req_workflow_acces			
28	req_04	MUST	#req_live_web_servic			
29	req_05	MUST	#req_inputs_specified			
30	req_06	MUST	#req_inputs_accessib			
31						
32	**Define rules to test individual requirements**					
33						
34	**Rule:**	#req_workflow_description				
35		**Exists:**	?wf rdf:type wfdesc:W ; rdfs:label ?label			
36		**Pass:**	Workflow description			
37		**Fail:**	Workflow description			
38						
39	**Rule:**	#req_workflow_aggregated				
40		**ForEach:**	?wf rdf:type wfdesc:W ; rdfs:label ?wflab ; wfdesc:hasWorkflo			
41		**ResultMod:**	ORDER BY ?wflab			
42		**Aggregates:**	{+wfdef}			
43		**Pass:**	All referenced workflo			
44		**Fail:**	Workflow definition %			
45		**None:**	No workflow definitior			
46						
47	**Rule:**	#req_workflow_accessible				
		ForEach:	?wf rdf:type wfdesc:W			

Sample workflow for KEGG decay analysis

The workflow in this pack is a simple procedure that uses a KEGG service to show in a browser pubmed articles related to an input pathway.

✗ **Target Pack390 *does not satisfy* checklist for *complete.***

- ✓ Workflow description metadata is present
- ✓ All referenced workflow definitions are aggregated
- ✓ All referenced workflow definitions are accessible
- ✗ Workflow service http://soap.genome.jp/KEGG.wsdl for get_references_by_pathway is not accessible
- ✗ No input data indicated for Retrieve_Pubmed_Publ
- ✓ All specified input files are accessible

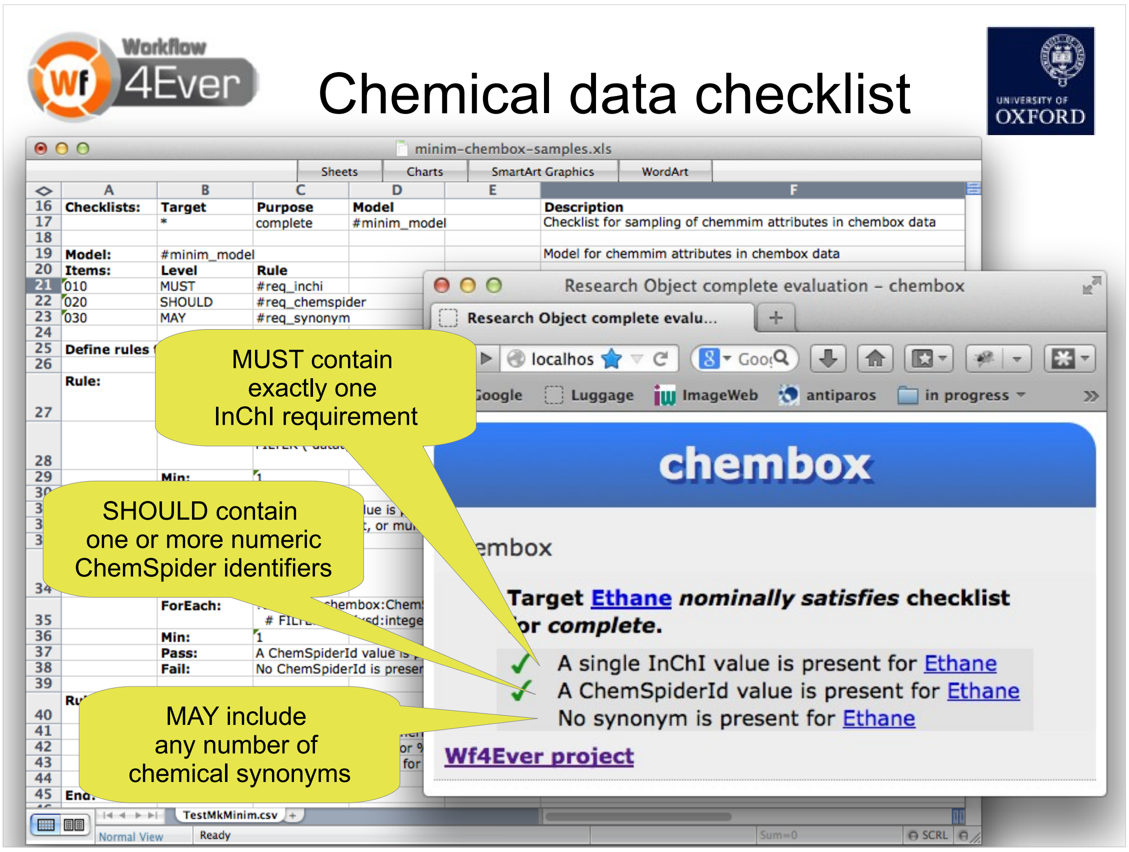
Wf4Ever project

@@SKIP

To illustrate these ideas in a concrete example...

One checklist we have worked with, which is intended to give an indication of whether a workflow is, prima facie, likely to be runnable, without having to gather the resources needed to actually run it.

To declare a workflow to be runnable, we require that:
- a workflow description is present in the RO
- the RO includes an executable workflow definition
- the workflow definition is accessible
- all services references by the workflow are accessible
- all inputs required by the workflow are defined
- all inputs are accessible



Chemical data checklist

minim-chembox-samples.xls

	A	B	C	D	E	F
16	**Checklists:**	**Target**	**Purpose**	**Model**		**Description**
17		*	complete	#minim_model		Checklist for sampling of chemmim attributes in chembox data
18						
19	**Model:**	#minim_model				Model for chemmim attributes in chembox data
20	**Items:**	**Level**	**Rule**			
21	010	MUST	#req_inchi			
22	020	SHOULD	#req_chemspider			
23	030	MAY	#req_synonym			

Define rules

Rule:

Min: 1

ForEach: chembox:Chem
FIL xsd:intege

Min: 1

Pass: A ChemSpiderId value is

Fail: No ChemSpiderId is preser

End

TestMkMinim.csv

Research Object complete evaluation – chembox

chembox

Target Ethane *nominally satisfies* checklist for *complete*.

✓ A single InChI value is present for Ethane
✓ A ChemSpiderId value is present for Ethane
No synonym is present for Ethane

Wf4Ever project

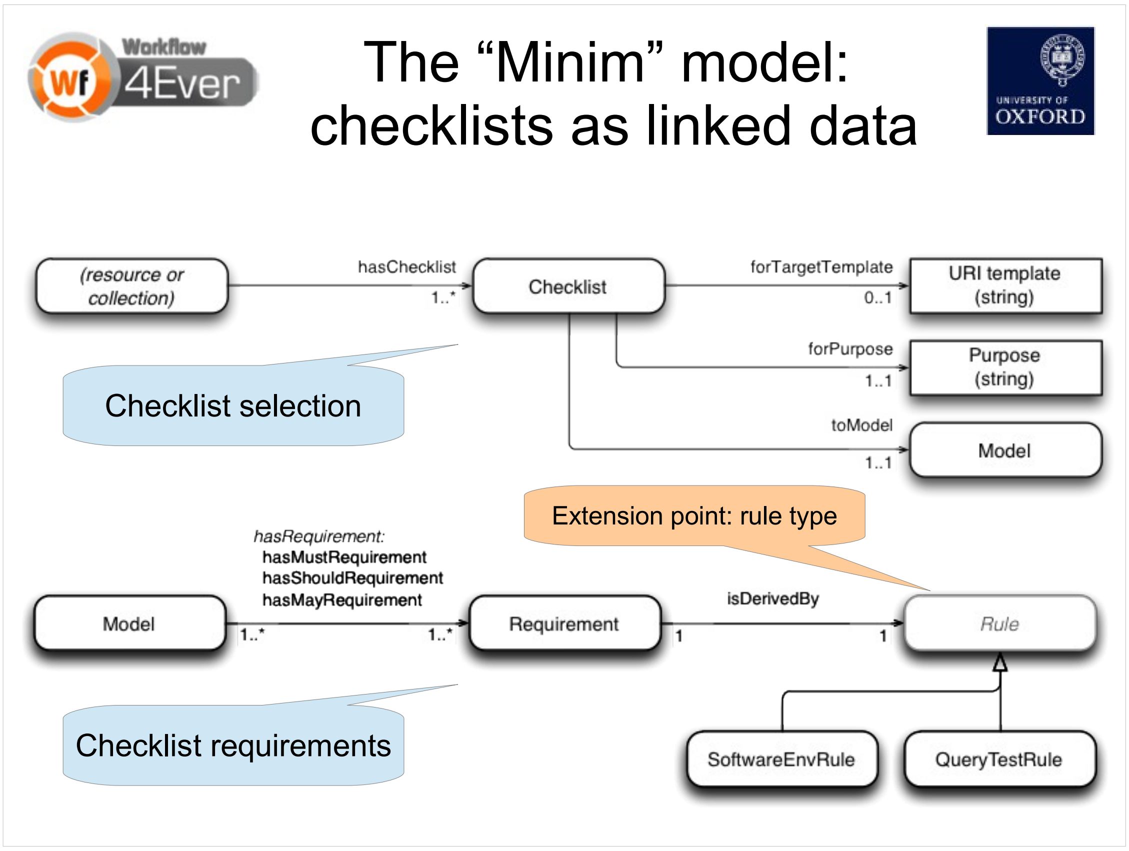
MUST contain exactly one InChI requirement

SHOULD contain one or more numeric ChemSpider identifiers

MAY include any number of chemical synonyms

To illustrate this idea of requirement levels in a concrete example, consider this evaluation of completeness of chemical information...

Among other things, community norms expect a that chemical description:
- MUST include exactly one International Chemical Identif er (or InChI), presented as a string
- SHOULD include at least one ChemSpider identifier, which is an integer value
- MAY include any number of synonyms



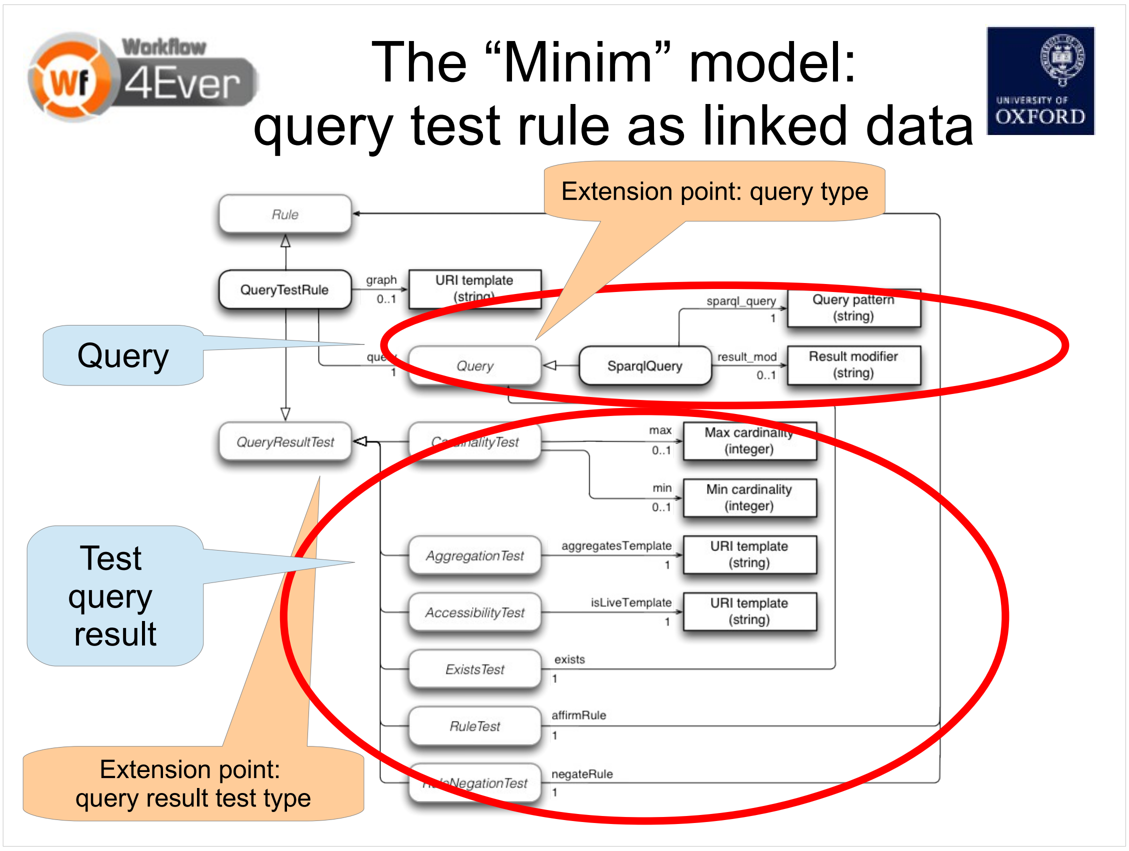
The “Minim” model: checklists as linked data

(resource or collection)
hasChecklist
1..*
Checklist
forTargetTemplate
0..1
URI template (string)
forPurpose
1..1
Purpose (string)
toModel
1..1
Model
Checklist selection
Extension point: rule type
hasRequirement:
hasMustRequirement
hasShouldRequirement
hasMayRequirement
Model
1..*
1..*
Requirement
isDerivedBy
1
1
Rule
SoftwareEnvRule
QueryTestRule
Checklist requirements

We have defined our “Minim” model to represent these checklists as linked data. I shalln't go through the model in detail, but I shall highlight some key features.

It uses three main sub-components: selectors, checklists and rules:
 - checklist selectors are used to select a particular checklist from those available based on a target resource, and the purpose for which it is evaluated (e.g. evaluating a specified chemical description for completeness, or a specified workflow for runnability)
 - the actual checklist, or Model, which is simply a list of requirements with associated requirement levels
 - rules that are used to evaluate the requirements (e.g. does a chemical description include an InChI string value?)

The evaluation rule represents an extension point in the model, where new rule types may be introduced as required. Here, just two rule types are shown. I shall focus on just one of these...



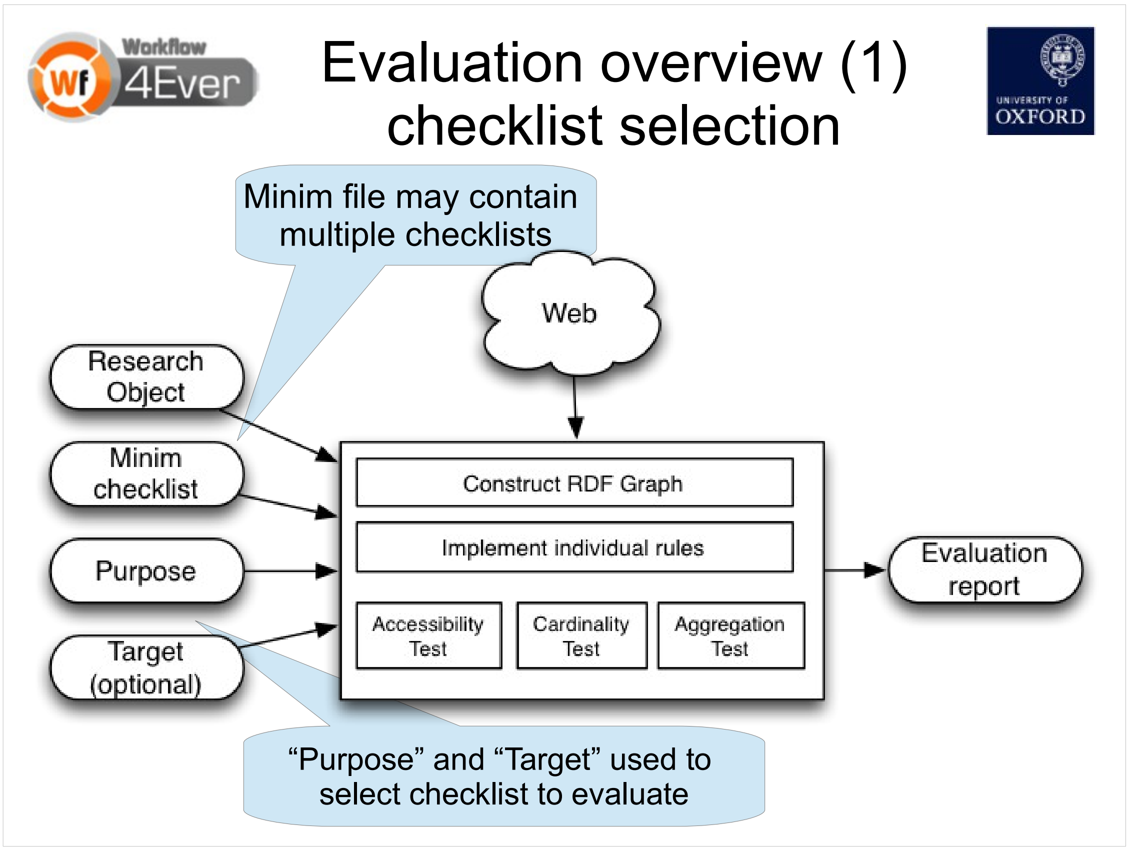
The “Minim” model:
query test rule as linked data

Currently, almost all checklist requirements are evaluated using a “query test rule”, which has two key elements:

 - a query that is evaluated against the combined RO annotations (e.g. to query for InChI identifiers), and
 - a test that is applied to the result of the query (e.g. is exactly one InChI identifier result returned?). Some tests may request additional information from the RO, or from the wider Web (e.g. to test if a resource is aggregated, or accessible)

Queries used are SPARQL graph patterns, such as appear in a SPARQL WHERE clause. But the query type is an extension point in the model where different formats may be introduced (e.g. queries based on SPIN, or OWL class expressions could be added here).

Query result tests include existence, cardinality, resource accessibility, etc. This, too, is an extension point where new features may be introduced.



Evaluation overview (1) checklist selection

Minim file may contain multiple checklists

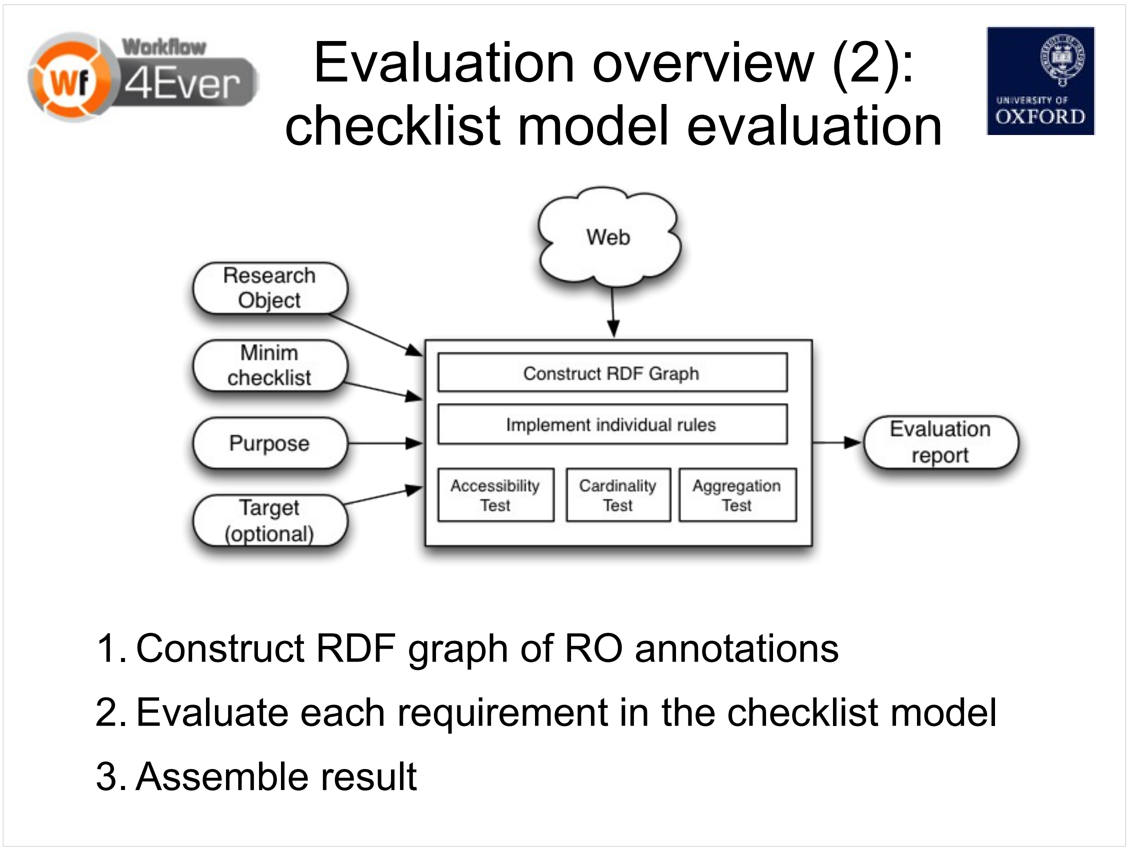
“Purpose” and “Target” used to select checklist to evaluate

The checklist evaluation proceeds by first selecting a checklist from those available, based upon supplied target resource and purpose values (e.g. is a specified chemical description complete, or is a specified workflow runnable?)

((Checklist evaluation of an RO uses four input values:

- a research object providing the evaluation context
- a Minim checklist resource defining one or more checklists
- an optional target resource (if not specified, the RO itself is the target resource). E.g. a particular workflow within the RO.
- a purpose identifier, which is just a string used to distinguish different purposes for which evaluation may be performed (e.g. “complete”, “runnable”, etc.)

The first step of checklist evaluation is to select a checklist from the Minim resource that matches the supplied target and purpose values.))



Evaluation overview (2): checklist model evaluation

1. Construct RDF graph of RO annotations
2. Evaluate each requirement in the checklist model
3. Assemble result

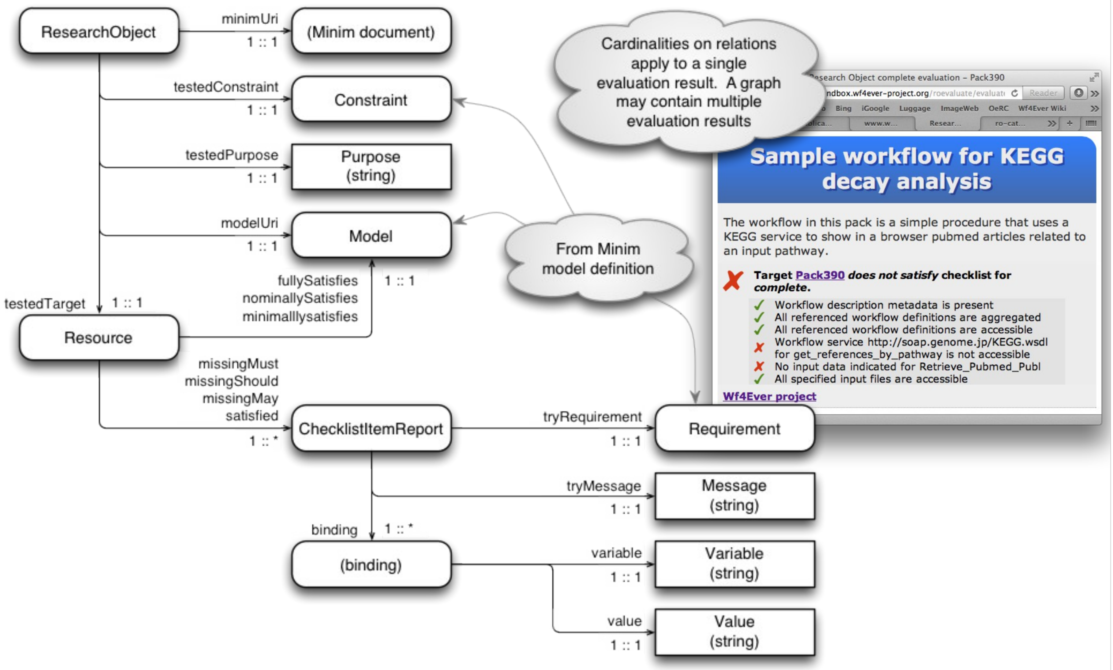
Having selected a checklist, the main evaluation can proceed:

1. An RDF graph is assembled from the RO annotations

2. Each requirement in the checklist is evaluated in the context of the assembled RDF graph, yielding a True or False result

3. A final result is assembled from the individual results, taking account of the corresponding requirement levels

Checklist evaluation: results as linked data



Sample workflow for KEGG decay analysis

The workflow in this pack is a simple procedure that uses a KEGG service to show in a browser pubmed articles related to an input pathway.

✗ **Target Pack390 *does not satisfy* checklist for *complete*.**

- ✓ Workflow description metadata is present
- ✓ All referenced workflow definitions are aggregated
- ✓ All referenced workflow definitions are accessible
- ✗ Workflow service http://soap.genome.jp/KEGG.wsdl for get_references_by_pathway is not accessible
- ✗ No input data indicated for Retrieve_Pubmed_Publ
- ✓ All specified input files are accessible

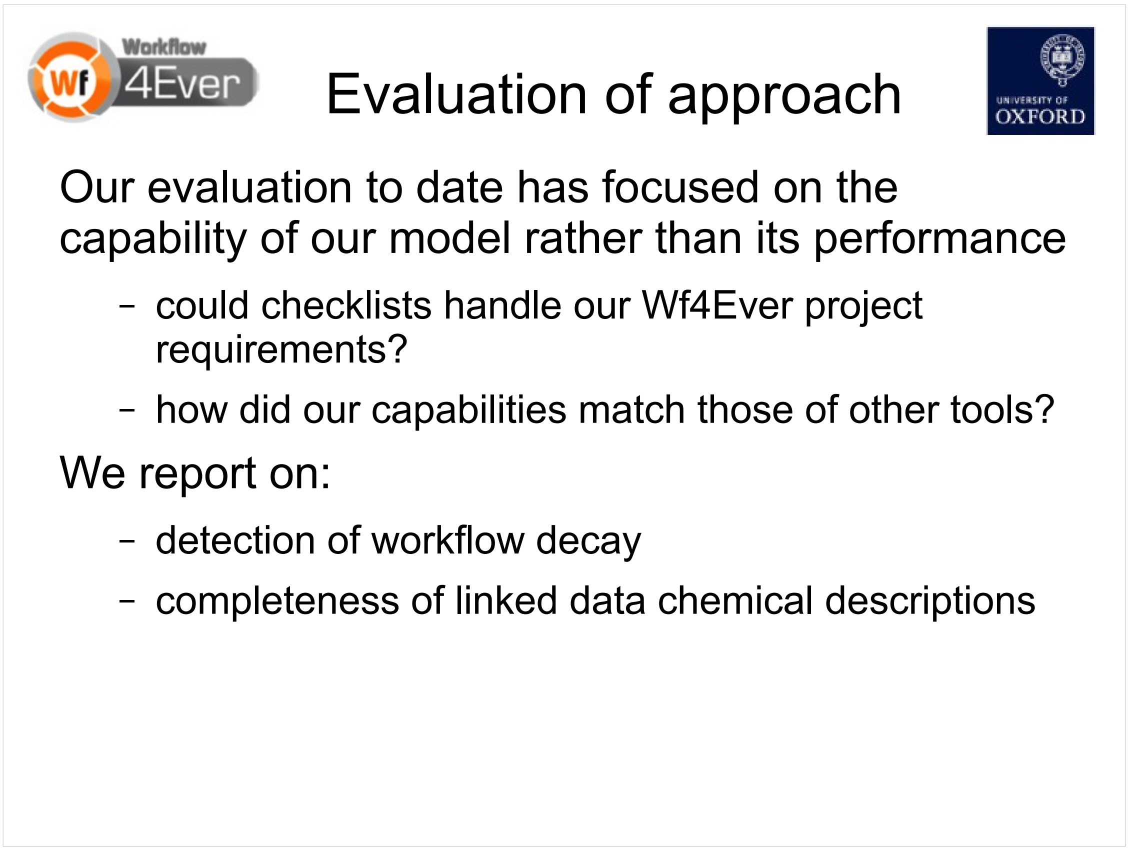
Wf4Ever project

The final result of a checklist evaluation also represented as an RDF graph, and could itself be exposed as linked data.

I'm not going to go into the details now, but it might be worth noting that the result graph links directly back to the evaluated target resource.

We have also implemented further services that process the RDF graph result to generate a more easily used “traffic light” summary of the evaluation, which is available as either HTML or JSON.

Evaluation and Reflections



Evaluation of approach

Our evaluation to date has focused on the capability of our model rather than its performance

- could checklists handle our Wf4Ever project requirements?
- how did our capabilities match those of other tools?

We report on:

- detection of workflow decay
- completeness of linked data chemical descriptions

Our evaluation has focused on capabilities of our tool rather than its performance; i.e. can we perform the evaluations required in the Wf4Ever project, and how do our capabilities compare with previous work?

We did not set out to evaluate performance, as we did not expect it to be an issue in the envisaged uses of checklists (but I'll return to this later.)

Our main capability evaluation has been performed with respect to two applications:

- one was workflow decay detection – which was a requirement arising from our own Wf4Ever project.

- and the other evaluation was completeness of chemical information - which was a comparison with some previous quality evaluation work

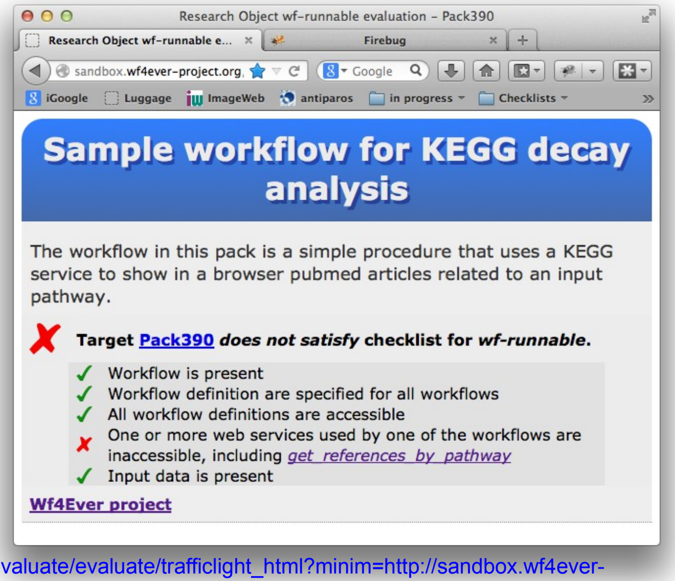
Workflow Decay Detection

2012: replacement of KEGG Web Services with REST services

Workflows located in myExperiment

Before shutdown, workflow runnability was confirmed

After shutdown, the checklist reports workflow decay


Research Object wf-runnable evaluation – Pack390
Research Object wf-runnable e...
Firebug
sandbox.wf4ever-project.org
Google
iGoogle
Luggage
ImageWeb
antiparos
in progress
Checklists

Sample workflow for KEGG decay analysis

The workflow in this pack is a simple procedure that uses a KEGG service to show in a browser pubmed articles related to an input pathway.

✗ Target Pack390 *does not satisfy* checklist for *wf-runnable*.

- ✓ Workflow is present
- ✓ Workflow definition are specified for all workflows
- ✓ All workflow definitions are accessible
- ✗ One or more web services used by one of the workflows are inaccessible, including *get_references_by_pathway*
- ✓ Input data is present

Wf4Ever project

http://sandbox.wf4ever-project.org/roevaluate/evaluate/trafficlight_html?minim=http://sandbox.wf4ever-project.org/rodl/ROs/Kegg-workflow-evaluation/Runnable-workflow-checklist.rdf&purpose=wf-runnable&RO=http://sandbox.wf4ever-project.org/rodl/ROs/Pack390/

Last year (2012), The Kyoto Encyclopedia of Genes and Genomes transitioned their web services to use a new REST interface

Before the old services were shut down, a number of client workflows were located in myExperiment, and run to confirm their viability

After the shutdown, the checklist service was run over these workflows, and was able to successfully detect and predict workflow decay caused by withdrawal of the web service

Completeness of Chemical Descriptions

Requirements testing with SPIN

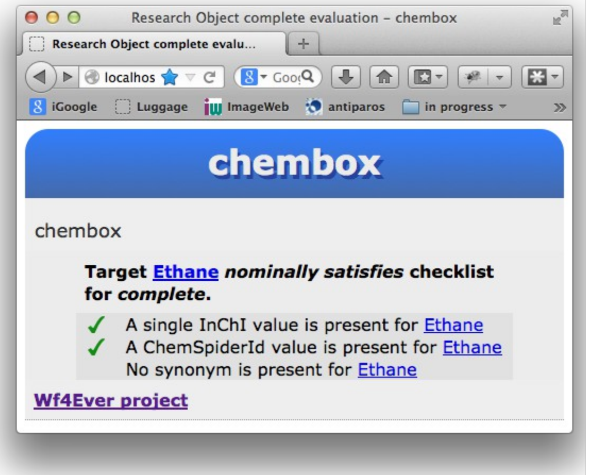
- Previous work, using SPARQL Inference Notation

"ChemBox" chemical descriptions extracted from Wikipedia

Key difference is query syntax:

- SPARQL vs SPIN

One test was unsuited to SPARQL query


Research Object complete evaluation – chembox
Research Object complete evalu...
localhos
Goo
iGoogle
Luggage
ImageWeb
antiparos
in progress

chembox

chembox

Target Ethane *nominally satisfies* checklist for *complete*.

✓ A single InChI value is present for Ethane
✓ A ChemSpiderId value is present for Ethane
No synonym is present for Ethane

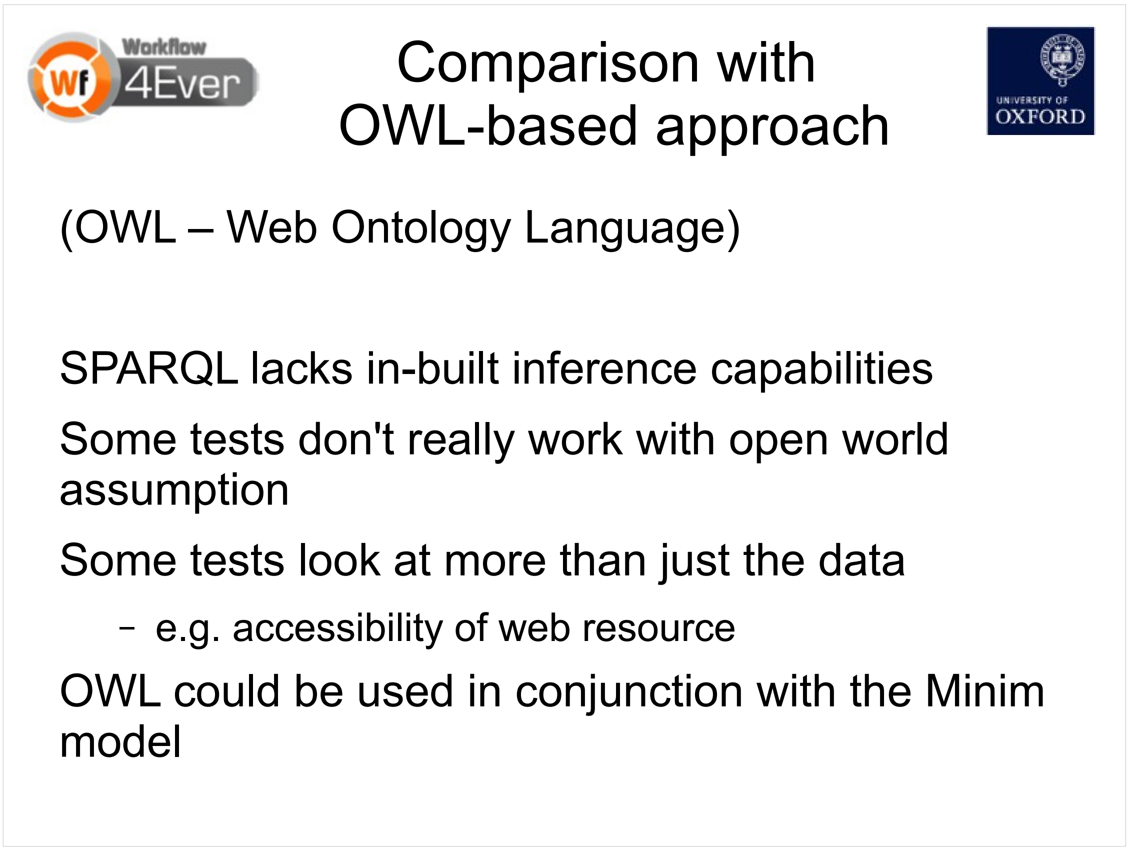
Wf4Ever project

The second application we evaluated targeted the completeness of "ChemBox" chemical descriptions extracted from Wikipedia info boxes.

The evaluation was a comparison with previous work by Matt Gamble that used SPIN as its data probing mechanism. We used our checklist tool to evaluate the same datasets that were used in Matt's work.

The outcome was that we could reproduce the results of the previous work, with the exception of one test involving a complex SPIN expression to match a chemical formula text, which I could not decode with the tools accessible to me.

Given than SPIN is primarily another way to represent SPARQL, it seems probable that this test could have been done using SPARQL. But the complexity of this query means that it may be unsuited for practical use. For this, I would probably choose to use the Minim extension points to apply a different approach, such as regex matching.



Comparison with OWL-based approach

(OWL – Web Ontology Language)

SPARQL lacks in-built inference capabilities

Some tests don't really work with open world assumption

Some tests look at more than just the data

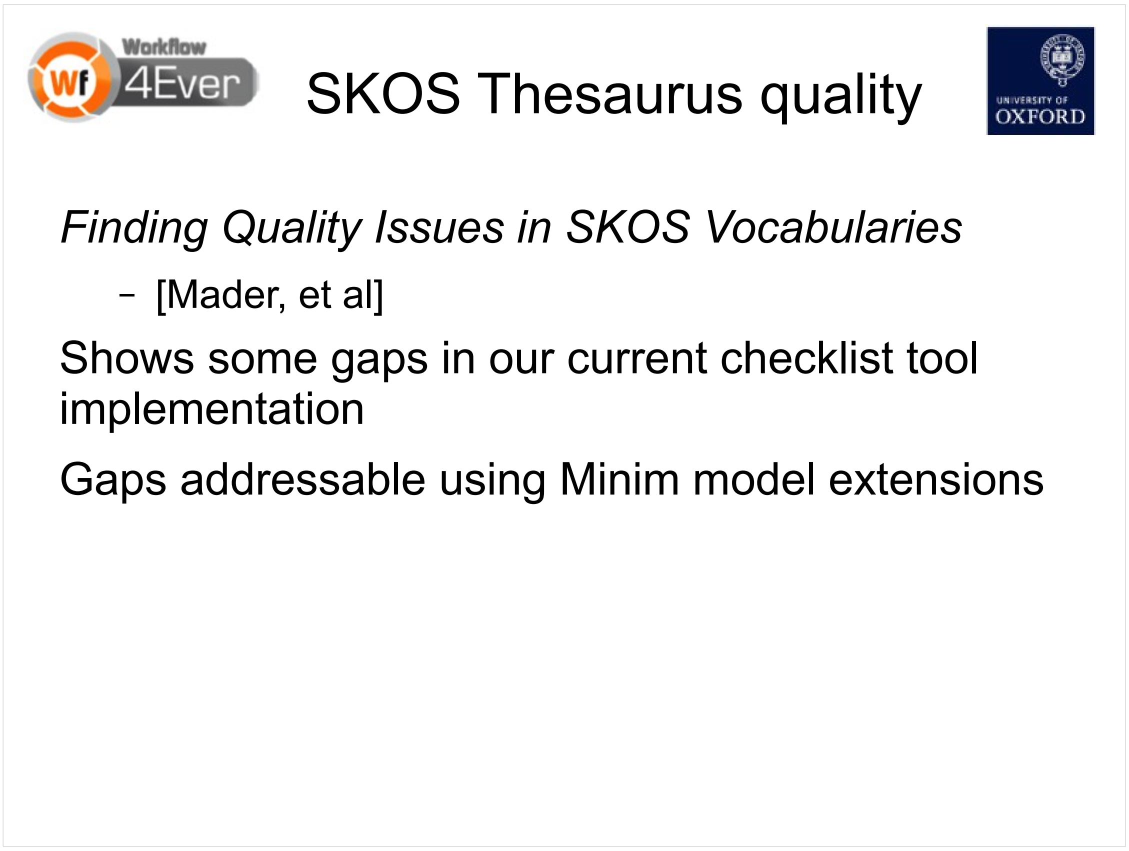
- e.g. accessibility of web resource

OWL could be used in conjunction with the Minim model

We also gave some consideration, but short of a full evaluation, to using OWL in our checklist model...

OWL inference provides an alternative way to generate requirement satisfaction reports, and is in some respects more expressive than SPARQL for this. But, some of our tests, such as cardinality tests, depend on treating an RO as a locally closed world, which is at odds with standard OWL satisfaction semantics.

Further, we use SPARQL queries in conjunction with additional tests that cannot be evaluated from the data alone (e.g. testing if a resource is accessible on the web). Our “Minim” model combines these elements (e.g. when checking for accessibility of workflow inputs). We could, in principle, use OWL instance retrieval as an alternative to SPARQL for probing the RO data.



SKOS Thesaurus quality

Finding Quality Issues in SKOS Vocabularies

- [Mader, et al]

Shows some gaps in our current checklist tool implementation

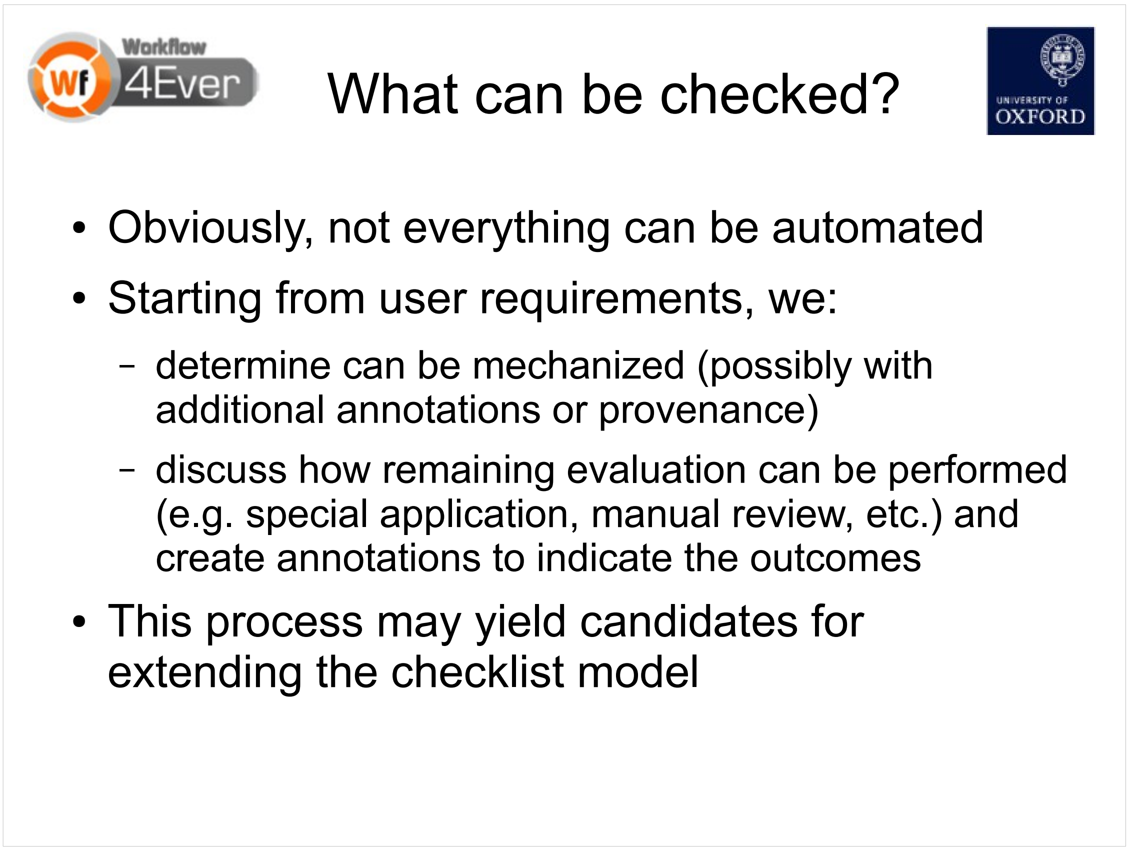
Gaps addressable using Minim model extensions

@@SKIP if < 5 mins

Separately, and not reported in our paper, we have also looked at some work on SKOS thesaurus quality evaluation [1], which has exposed some gaps in our current capabilities.

These gaps appear to be addressable using a small number of extensions to the Minim model

[1] Finding Quality Issues in SKOS Vocabularies
Christian Mader, Bernhard Haslhofer, and Antoine Isaac



What can be checked?

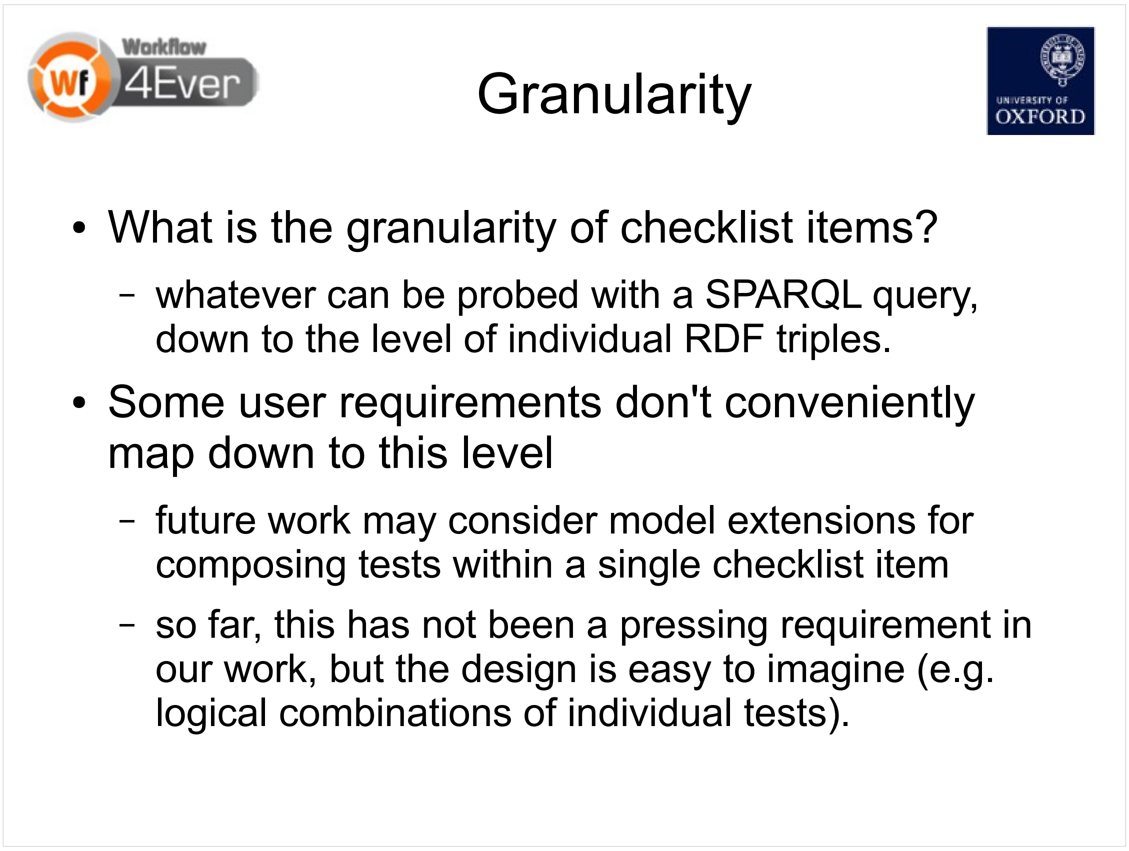
- Obviously, not everything can be automated
- Starting from user requirements, we:
  - determine can be mechanized (possibly with additional annotations or provenance)
  - discuss how remaining evaluation can be performed (e.g. special application, manual review, etc.) and create annotations to indicate the outcomes
- This process may yield candidates for extending the checklist model

Some questions that have arisen in reviews of our paper are:
“what can be checked by our tool” and
“what can be checked automatically?”

Obviously, not every aspect of quality can be checked automatically. Our approach has been to start from user requirements and divide them into those that can easily be handled as automatic tests and others that require more special treatment.

The cases requiring special treatment, such as manual review, are assumed to be handled separately, and then adding annotations to the RO indicating that the required review, or evaluation, has been performed. Tests for these annotations can then be included in a checklist.

These special cases may, in turn, suggest candidates for new automated testing capabilities.



Granularity

- What is the granularity of checklist items?
  - whatever can be probed with a SPARQL query, down to the level of individual RDF triples.
- Some user requirements don't conveniently map down to this level
  - future work may consider model extensions for composing tests within a single checklist item
  - so far, this has not been a pressing requirement in our work, but the design is easy to imagine (e.g. logical combinations of individual tests).

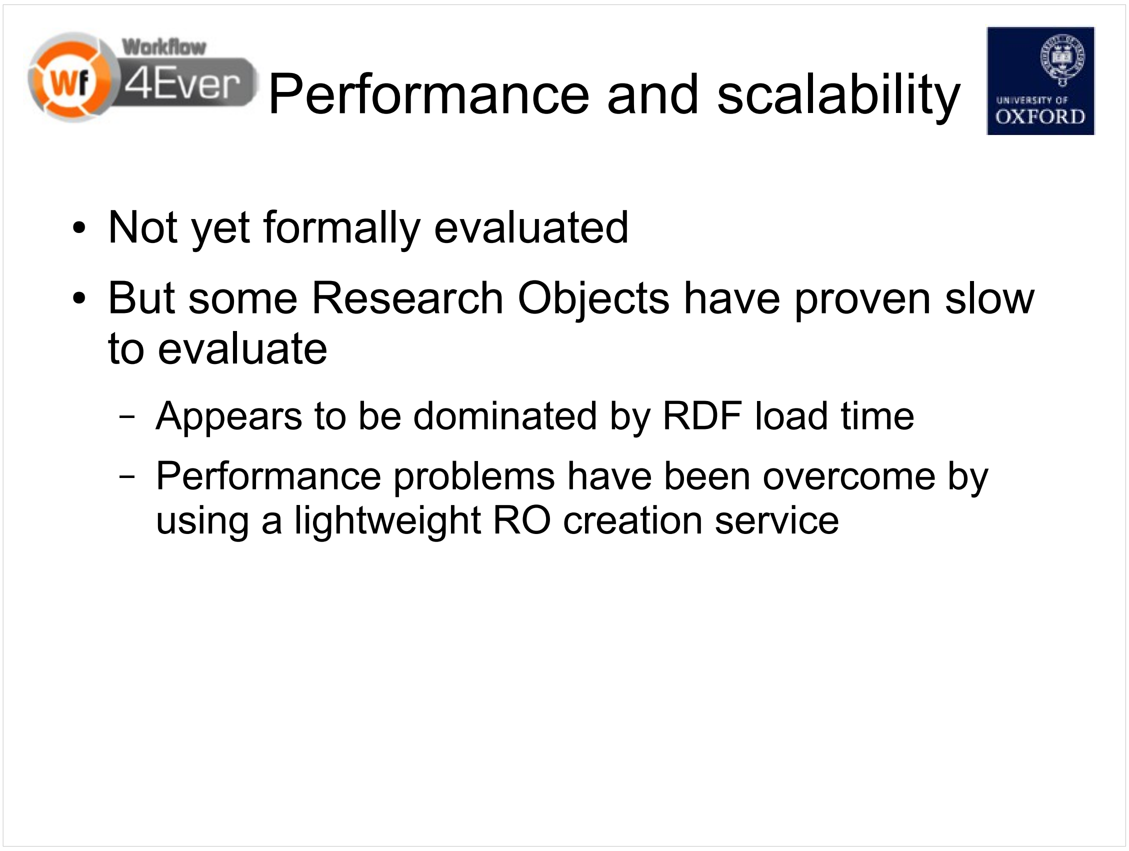
@@Skip slide if <4 mins to go

Another question that has been asked is “what is the granularity of checklist items?”

A short answer is: whatever granularity can be probed by a SPARQL query. This means that granularity can be down to the level of individual RDF triples in the RO annotations.

But we have noticed that some user requirements don't necessarily match the granularity that is conveniently offered by SPARQL, and future work may consider model extensions for composing multiple tests in a checklist item.

So far, this has not been an issue for us, but it is easy to imagine extension structures that provide for logical combinations of existing tests.



Performance and scalability

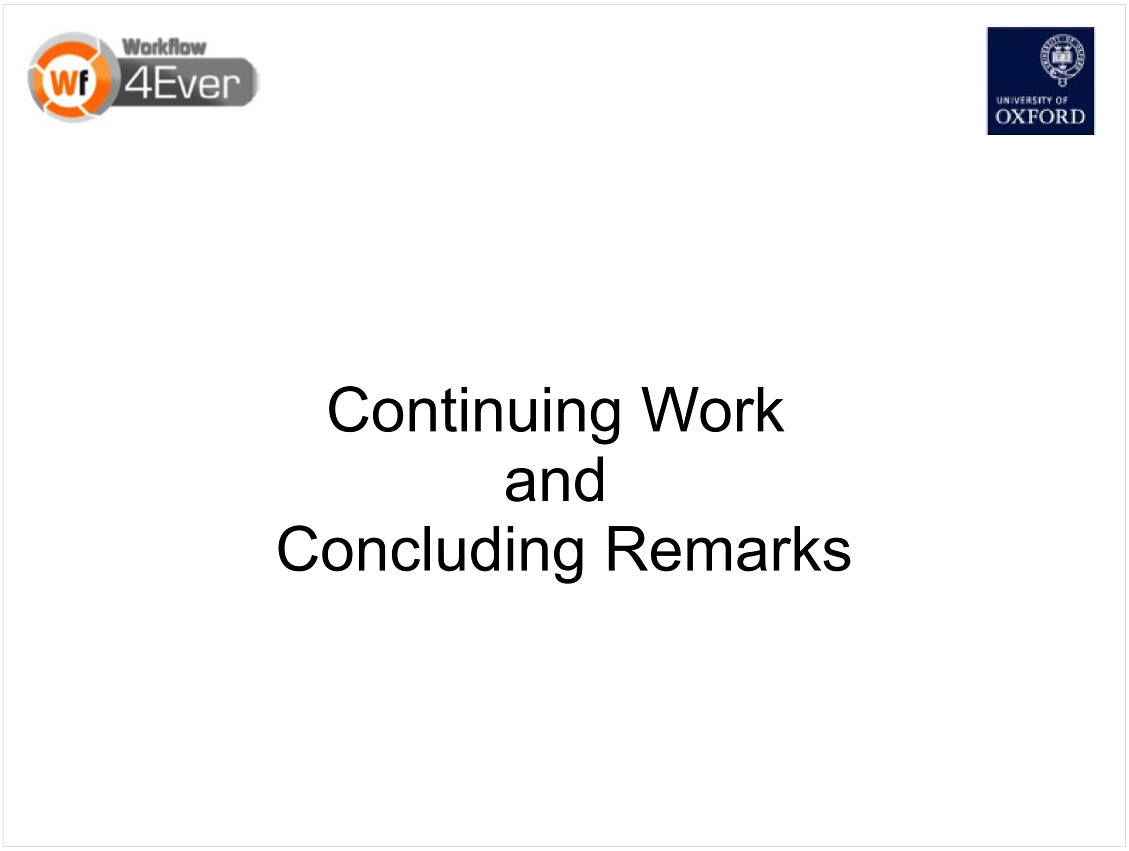
- Not yet formally evaluated
- But some Research Objects have proven slow to evaluate
  - Appears to be dominated by RDF load time
  - Performance problems have been overcome by using a lightweight RO creation service

As noted previously, we did not expect performance to be a concern for the envisaged usage scenarios, and did not undertake a formal evaluation of speed and scalability of our tool.

But we did run into some performance issues, notably in checking completeness of chemical information, where our initial attempts used a single RO with data about some 7500 chemicals.

In this, and in other instances where we have seen performance issues, they were due to RDF loading times rather than the checklist evaluation itself.

For the chemical evaluation, we have since resolved the performance problem by using a lightweight RO creation service (which is mentioned later).

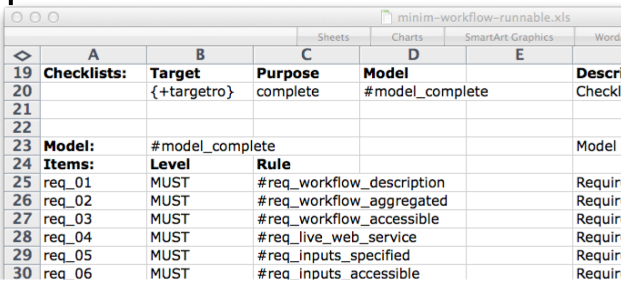


Continuing Work and Concluding Remarks

Finally, I'll mention some ongoing and possible future work, and offer some concluding remarks.

Recent and ongoing work

Minim creation from spreadsheet



minim-workflow-runnable.xls

◇	A	B	C	D	E	
19	**Checklists:**	**Target**	**Purpose**	**Model**		**Descri**
20		{+targetro}	complete	#model_complete		Checkl
21						
22						
23	**Model:**	#model_complete				Model
24	**Items:**	**Level**	**Rule**			
25	req_01	MUST	#req_workflow_description			Requir
26	req_02	MUST	#req_workflow_aggregated			Requir
27	req_03	MUST	#req_workflow_accessible			Requir
28	req_04	MUST	#req_live_web_service			Requir
29	req_05	MUST	#req_inputs_specified			Requir
30	req_06	MUST	#req_inputs_accessible			Requir

Evaluating arbitrary linked data

- "Overlay RO" service
- lightweight ROs for linked data

Matching/aligning quality metrics with checklist capabilities

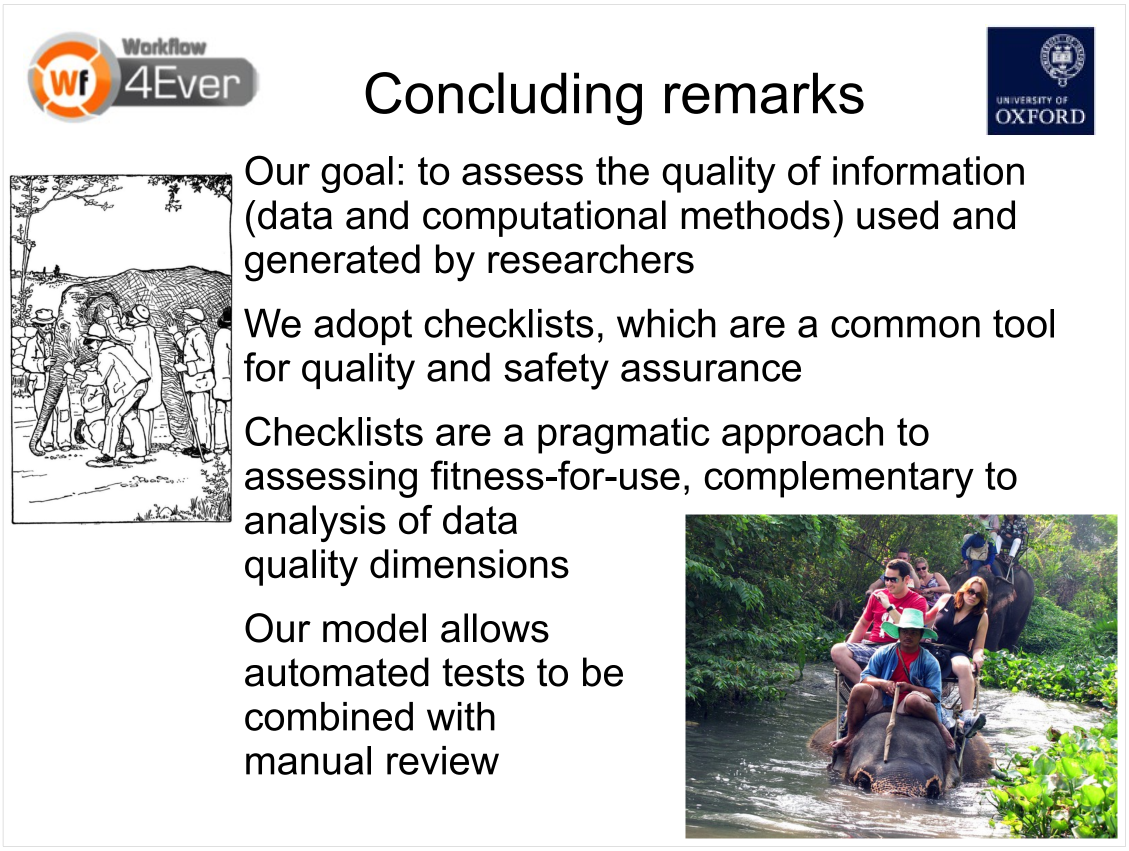
Checklist catalogue

Until recently, we created Minim checklists by hand editing RDF, but this is clearly not a viable solution for most users. We have since created a tool which uses a spreadsheet as the original checklist source and converts it to a Minim description in RDF. We also have some other ideas for more approachable tools for authoring checklist descriptions.

Further, we aim to apply checklists to any linked data (not necessarily supplied in an RO). To this end, we are experimenting with a lightweight "Overlay RO" service that allows linked data to be presented as an RO for evaluation.

Using this Overlay RO service, we were able to overcome the Chembox evaluation performance bottleneck that is mentioned in our paper, and alluded to earlier.

Other work under consideration includes aligning work on quality metrics and dimensions with checklist evaluation, and creation of a checklist repository for common evaluation requirements.



Concluding remarks

Our goal: to assess the quality of information (data and computational methods) used and generated by researchers

We adopt checklists, which are a common tool for quality and safety assurance

Checklists are a pragmatic approach to assessing fitness-for-use, complementary to analysis of data quality dimensions

Our model allows automated tests to be combined with manual review

In summary, our goal has been to assess the quality of information, including both data and computational artifacts, that informatics-based researchers build upon in their work. Thus, for example, we hope to contribute to enhancing the reproducibility of *in silico* research.

Our adoption of checklists is a pragmatic approach to fitness-for-use evaluation of scientific information, which is complementary to existing work on quality dimension analysis, and associated quality metrics

The checklist model is flexible, extensible and allows automated tests to be combined with manual review and other processes to provide a comprehensive coverage of quality evaluation requirements

Acknowledgements

EU Wf4Ever project (270129)
funded under EU FP7 (ICT- 2009.4.1).
(http://www.wf4ever-project.org)

Finally, we thank the many other researchers in the Wf4Ever project, who have provided valuable motivation and criticisms that have helped to guide the direction of this work.

Links

- Paper
  - …
- Presentation
  - …
- Software
  - https://github.com/wf4ever/ro-manager
- Evaluation scripts and data
  - https://github.com/wf4ever/ro-catalogue/tree/master/v0.1/minim-evaluation