Overview of Provenance on the Web

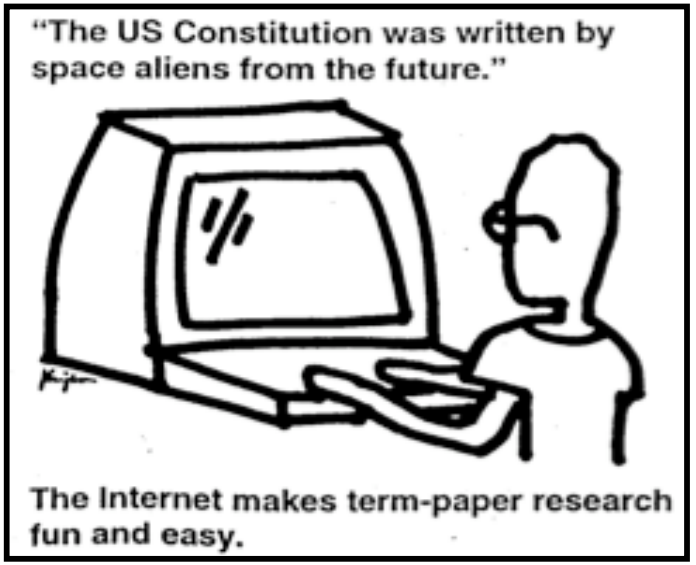
by the

W3C Provenance Incubator Group

Semantic Web Activity
World Wide Web Consortium

http://www.w3.org/2005/Incubator/prov/wiki

Special thanks to contributing group members: Chris Bizer, James Cheney, Sam Coppens, Kai Eckert, Andre Freitas, Irini Fundulaki, Daniel Garijo, Yolanda Gil, Jose Manuel Gomez Perez, Paul Groth, Olaf Hartig, Deborah McGuinness, Simon Miles, Paolo Missier, Luc Moreau, James Myers, Michael Panzer, Paulo Pinheiro da Silva, Christine Runnegar, Satya Sahoo, Yogesh Simmhan, Raphaël Troncy, Jun Zhao



Outline

- Need for provenance on the web
- W3C Provenance Group: What is known about provenance in the community
  - Definition of provenance
  - Key dimensions of provenance
  - Use cases: Requirements for provenance
  - State-of-the-art and existing provenance vocabularies
  - Situating provenance on the Web architecture
- Towards a standard for provenance

Provenance is Key in Open Information Systems (such as the Web)

- Provenance questions useful for information integration:
  - Who created that content (author/attribution)?
  - Was the content ever manipulated, if so by what processes/entities?
  - Who is providing that content (repository)?
  - What is the timeliness of that content?
  - Can any of the answers to these questions be verified (eg e-signatures)?

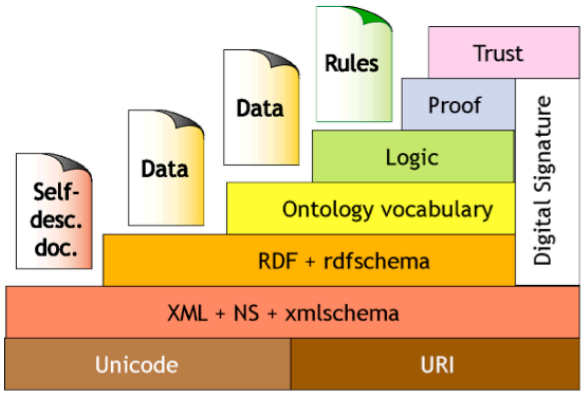
Broad Need for Provenance in Many Areas

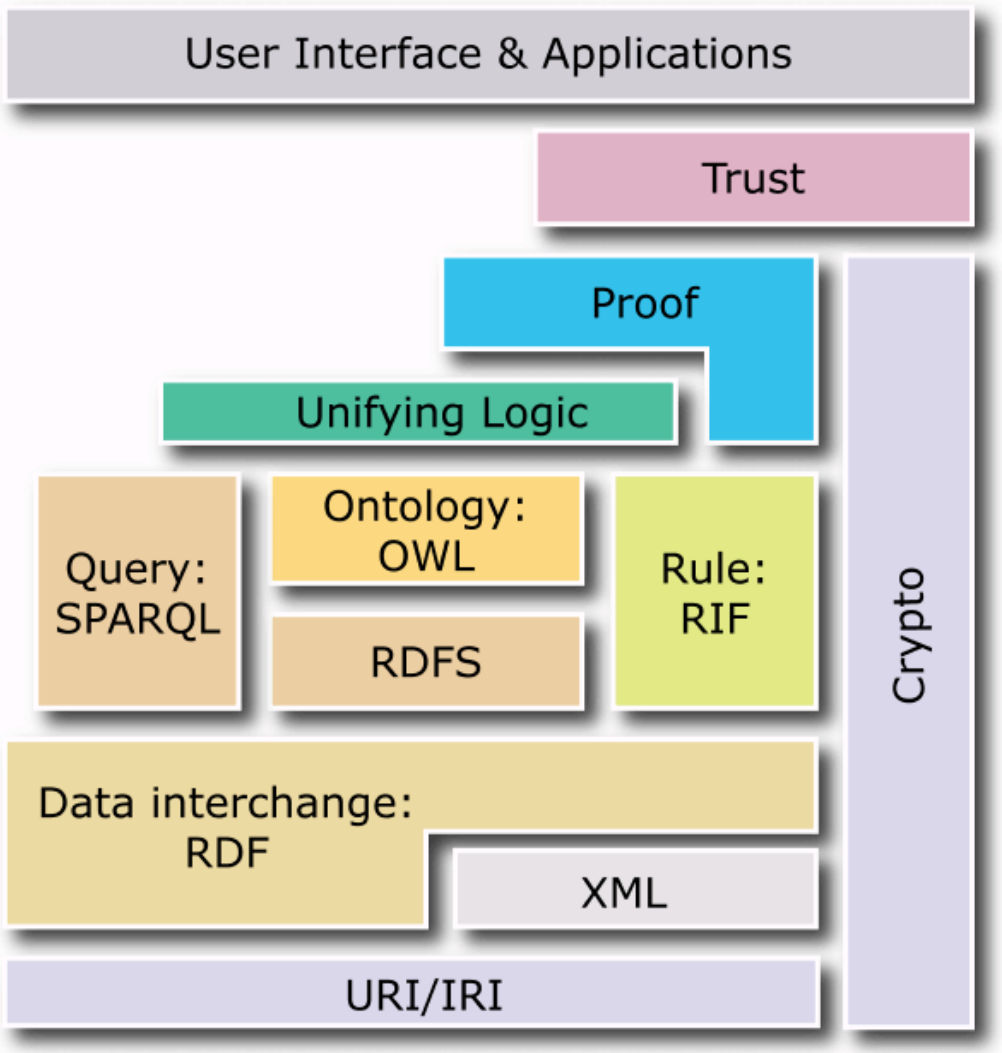
- Open information systems (such as the Web)
  - Making trust judgments on what web content to trust
- Business practices
  - Manufacturing processes and providers of a given product
- Science applications
  - How new results were obtained: from assumptions to conclusions and everything in between
- News-spheres
  - Blogosphere, twittosphere
- Laws for IP and privacy protection
  - Licensing and attribution of a document/software that combines permissions and rights of text, images, etc.
  - Privacy of information as well as of its provenance

Provenance and “Web Design Issues”

"At the toolbar (menu, whatever) associated with a document there is a button marked "Oh, yeah?". You press it when you lose that feeling of trust. It says to the Web, 'so how do I know I can trust this information?'. The software then goes directly or indirectly back to metainformation about the document, which suggests a number of reasons.”

- T. Berners-Lee, Web Design Issues, September 1997





Provenance in Web Documents, Blogosphere

"The problem is - and this is true of books and every other medium - we don't know whether the information we find [on the Web] is accurate or not. We don't necessarily know what its provenance is." – Vint Cerf

"In content, as creation becomes overabundant and as value shifts from creator to curator, it becomes all the more vital to properly cite and link to sources [...]. Good curation demands good provenance. [...] Provenance is no longer merely the nicety of artists, academics, and wine makers. It is an ethic we expect." – Jeff Jarvis

- Illustrates the need for provenance for attribution, licensing, making trust judgments

Provenance in Open Government

"**Provenance is the number one issue** that we face when publishing government data in data.gov.uk" -- John Sheridan, UK National Archives, data.gov.uk

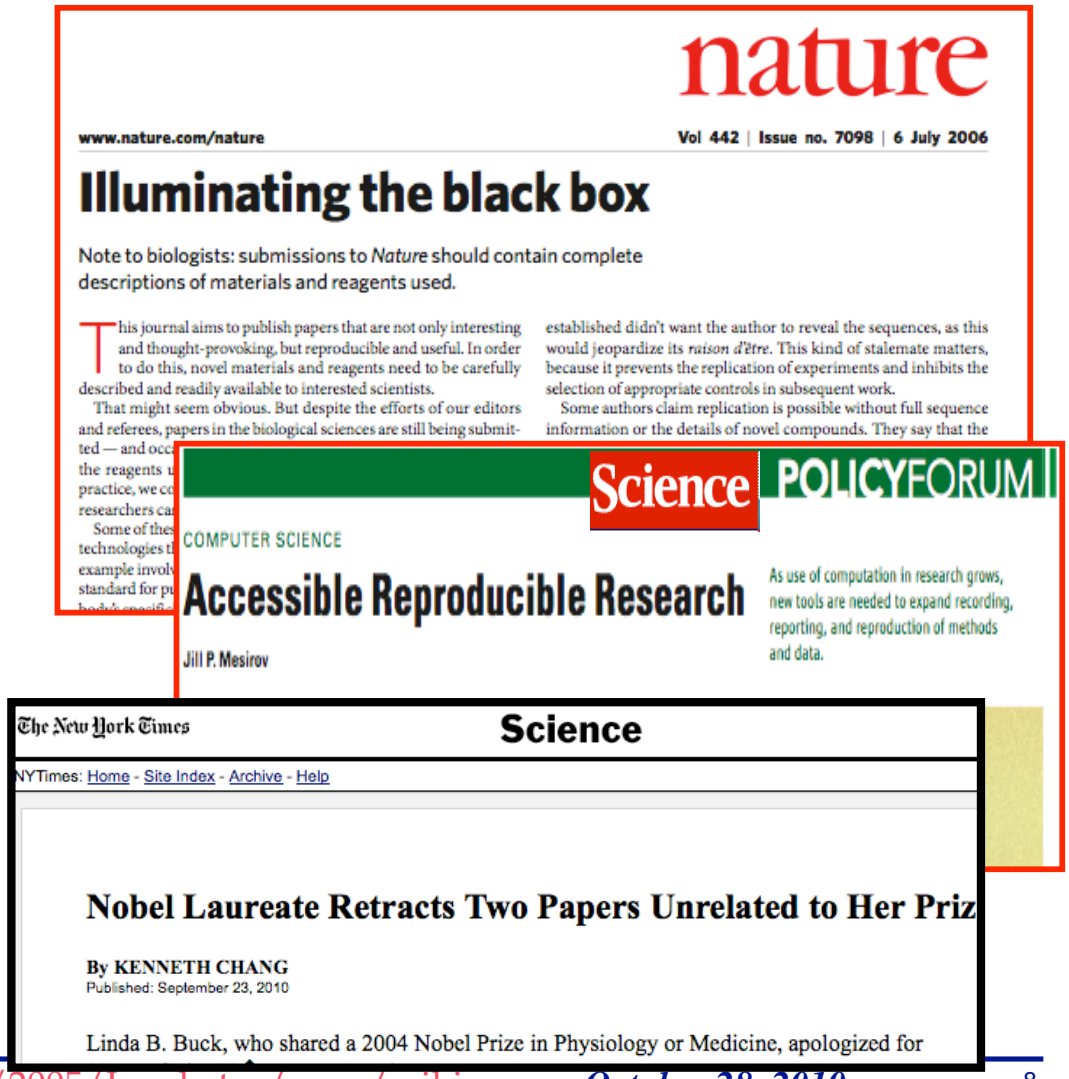
- Illustrates the need for provenance for data integration and reuse
  - Sources of data are very diverse
  - Varying quality
  - Different scope
  - Different assumptions

Provenance in Science

"We need a paradigm that makes it simple [...] to perform and publish reproducible computational research. [...] A Reproducible Research Environment (RRE) [...] provides computational tools together with the ability to automatically track the provenance of data, analyses, and results and to package them (or pointers to persistent versions of them) for redistribution."

- Jill Mesirov, Chief Informatics Officer of the MIT/Harvard Broad Institute, in Science, January 2010

- Illustrates the need for provenance for reproducibility and verification of processes



nature

www.nature.com/nature Vol 442 | Issue no. 7098 | 6 July 2006

Illuminating the black box

Note to biologists: submissions to *Nature* should contain complete descriptions of materials and reagents used.

This journal aims to publish papers that are not only interesting and thought-provoking, but reproducible and useful. In order to do this, novel materials and reagents need to be carefully described and readily available to interested scientists.

That might seem obvious. But despite the efforts of our editors and referees, papers in the biological sciences are still being submitted — and occ... the reagents u... practice, we co... researchers ca... Some of thes... technologies t... example involv... standard for pu...

established didn't want the author to reveal the sequences, as this would jeopardize its *raison d'être*. This kind of stalemate matters, because it prevents the replication of experiments and inhibits the selection of appropriate controls in subsequent work.

Some authors claim replication is possible without full sequence information or the details of novel compounds. They say that the

Science POLICYFORUM

COMPUTER SCIENCE

Accessible Reproducible Research

Jill P. Mesirov

As use of computation in research grows, new tools are needed to expand recording, reporting, and reproduction of methods and data.

The New York Times Science

NYTimes: Home - Site Index - Archive - Help

Nobel Laureate Retracts Two Papers Unrelated to Her Priz

By KENNETH CHANG
Published: September 23, 2010

Linda B. Buck, who shared a 2004 Nobel Prize in Physiology or Medicine, apologized for

Outline

- Need for provenance on the web
- W3C Provenance Group: What is known about provenance in the community
  - Definition of provenance
  - Key dimensions of provenance
  - Use cases: Requirements for provenance
  - State-of-the-art and existing provenance vocabularies
  - Situating provenance on the Web architecture
- Towards a standard for provenance

W3C Chartered a New Provenance Group in Sept'09 (Chair: Y. Gil)

- Provenance is a pressing issue in many areas for W3C
  - Linked Data and the semantic web (linkedopendata.org)
  - Open government (data.gov, data.gov.uk)
  - HCLS
- Most people do not know how to approach provenance
  - Many are asking for a standard and methodology that they can use immediately
- Existing work scattered in many areas of computer science and library sciences research
  - *"The number of publications on provenance is [...] a total of 425 [...] with about half the papers published in the last two years." – Luc Moreau*

http://www.w3.org/2005/Incubator/prov/charter

W3C Provenance Group: Original Charter

Provide state-of-the-art understanding and develop a roadmap for development and possible standardization

- Articulate requirements for accessing and reasoning about provenance information on an open system like the Web
  - Develop use cases
- Relate issues in provenance to Web architecture
  - Semantic Web, security, identity, etc.
- Report on state-of-the-art work on provenance
- Propose on a roadmap for provenance in the Semantic Web
  - Identify starting points for provenance representations
  - Identifying elements of a provenance architecture that would benefit from standardization

Outline

- Need for provenance on the web
- W3C Provenance Group: What is known about provenance in the community
  - Definition of provenance
  - Key dimensions of provenance
  - Use cases: Requirements for provenance
  - State-of-the-art and existing provenance vocabularies
  - Situating provenance on the Web architecture
- Towards a standard for provenance

Products of the W3C Provenance Group to Date

- Shared working **definition** of provenance (10/10)
- Developed a set of **key dimensions** for provenance (11/09)
  - Grouped into three major categories: content, management, use
- Collected **use cases** for provenance (12/09)
  - More than 30 use cases, most were improved and curated
- Designed 3 **flagship scenarios** from the use cases (4/10)
- Developed **provenance requirements** from the scenarios (6/10)
  - User requirements: what is the purpose/use of the provenance information
  - Technical requirements: derived from the user requirements
- Created **mappings** for existing provenance vocabularies (7/10)
- **State-of-the-art** report (9/10)
  - Need standards for publishing and accessing provenance
- Provenance in Web architecture (expected 11/15)
- Currently preparing recommendations and final report (expected 11/30)
  - Preparing a proposal charter for a working group on provenance

Our Working Definition of Provenance

Provenance of a resource is **a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource**.

Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility.

- Is provenance = metadata? Or = trust? Or = authentication?
  - Provenance can be seen as metadata, but not all metadata is provenance
  - Provenance provides a substrate for deriving different trust metrics
  - Provenance records can be used to verify and authenticate among other uses

- Notice:
  - Provenance assertions can have their own provenance
  - Inference is useful if provenance records are incomplete/erroneous
  - There may be alternative accounts of provenance of the same resource

Outline

- Need for provenance on the web
- W3C Provenance Group: What is known about provenance in the community
  - Definition of provenance
  - Key dimensions of provenance
  - Use cases: Requirements for provenance
  - State-of-the-art and existing provenance vocabularies
  - Situating provenance on the Web architecture
- Towards a standard for provenance

http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Dimensions

W3C Provenance Group: Major Dimensions of Provenance

1) Content

- Attribution - provenance as the sources or entities that were used to create a new result
  - Responsibility - knowing who endorses a particular piece of information or result
  - Origin - recorded vs reconstructed, verified vs non-verified, asserted vs inferred
- Process - provenance as the process that yielded an artifact
  - Reproducibility (eg workflows, mashups, text extraction)
  - Data Access (e.g. access time, accessed server, party responsible for accessed server)
- Evolution and versioning
  - Republishing (e.g. retweeting, reblogging, republishing)
  - Updates (eg a document with content from various sources and that changes over time)
- Justification for decisions – Includes argumentation, hypotheses, why-not questions
- Entailment - given the results to a particular query, what axioms or tuples led to those results

2) Management

- Publication - Making provenance information available (expose, distribute)
- Access - Finding and querying provenance information
- Dissemination control – Track policies specified by creator for when/how an artifact can be used
  - Access Control - incorporate access control policies to access provenance information
  - Licensing - stating what rights the object creators and users have based on provenance
  - Law enforcement (eg enforcing privacy policies on the use of personal information)
- Scale - how to operate with large amounts of provenance information

W3C Provenance Group: Major Dimensions of Provenance (Cont'd)

3) Use

- Understanding - End user consumption of provenance.
  - abstraction, multiple levels of description, summary
  - presentation, visualization
- Interoperability - combining provenance produced by multiple different systems
- Comparison - finding what's in common in the provenance of two or more entities (eg two experimental results)
- Accountability - the ability to check the provenance of an object with respect to some expectation
  - Verification - of a set of requirements
  - Compliance - with a set of policies
- Trust - making trust judgments based on provenance
  - Information quality - choosing among competing evidence from diverse sources (eg linked data use cases)
  - Incorporating reputation and reliability ratings with attribution information
- Imperfections - reasoning about provenance information that is not complete or correct
  - Incomplete provenance
  - Uncertain/probabilistic provenance
  - Erroneous provenance
  - Fraudulent provenance
- Debugging

Outline

- Need for provenance on the web
- W3C Provenance Group: What is known about provenance in the community
  - Definition of provenance
  - Key dimensions of provenance
  - Use cases: Requirements for provenance
  - State-of-the-art and existing provenance vocabularies
  - Situating provenance on the Web architecture
- Towards a standard for provenance

http://www.w3.org/2005/Incubator/prov/wiki/Use_Cases

W3C Provenance Group: 30+ Use Cases Contributed by the Community

1. Result Differences
2. Anonymous Information
3. Information Quality Assessment for Linked Data
4. Timeliness
5. Simple Trustworthiness Assessment
6. Ignoring Unreliable Data
7. Answering user queries that require semantically annotated provenance
8. Provenance in Biomedicine
9. Closure of Experimental Metadata
10. Locating Biospecimens With Sufficient Quality
11. Using process provenance for assessing the quality of Information products
12. Provenance Tracking in the Blogosphere
13. Provenance of a Tweet
14. Provenance and Private Data Use
15. Provenance of Decision Making in Emergency Response
16. Provenance of Collections vs Objects in Cultural Heritage
17. Provenance at different levels in Cultural Heritage
18. Identifying attribution and associations
19. Determining Compliance with a License
20. Documenting axiom formulation
21. Evidence for public policy
22. Evidence for engineering design
23. Fulfilling Contractual Obligations
24. Attribution for a versioned document
25. Provenance for Environmental Marine Data
26. Crosswalk Maintenance
27. Metadata Merging
28. Mapping Digital Rights
29. Computer Assisted Research
30. Handling Scientific Measurement Anomaly
31. Human-Executed Processes
32. Semantic disambiguation of data provider identity
33. Hidden Bug

Structure of the Use Cases

Owner

Chris Bizer
(Curator: Satya Sahoo)

Provenance Dimensions

Primary: Attribution (Content), Evolution and versioning (Content)
Secondary: Scale (Management), Law Enforcement (Management), Understanding (Use), Trust (Use), Incomplete provenance (Use)

Background and Current Practice

Within the Blogosphere, topics are discussed across blogs that refer to each other, for example on personal blogs, project weblogs, and on company blogs. The cross references are in the form of links at the bottom of a blog post, hyperlinks within a blog post, and quotation of text from other blogs. Blog posts are also aggregated and republished by services like Technorati, BlogPulse, Tailrank, and BlogScope, that track the interconnections between bloggers. Correct attribution of blogs, as they are processed, aggregated and republished on the Web, is an important requirement in the blogosphere.

Goal

Enable applications on the Web to attribute content from different sources to a specific individual or an organization. In this use case, blogs are an example of content flow between websites, and it is important to trace back republished posts to their original source.

Use Case Scenario

A website X collates Web content from multiple sites on a particular topic that is processed and aggregated for use by its customers. It is imperative for website X to present only credible and trusted content on its site to satisfy its customer, attract new business, and avoid legal issues. In the context of this blogosphere use case, a blog aggregator service or an user wants to identify the author of a blog without violating privacy laws. In some scenarios, the aggregator service or user may have only incomplete attribution information. In case the author of a blog is listed by name (first name, last name), disambiguation of an author is difficult with multiple blog authors sharing the same name and this may require use of additional user information (for example, email address) without violation of user privacy or privacy laws.

Problems and Limitations

The provenance of Web content in general and blog posts in particular are necessary to users for correct attribution and to aggregating services. Aggregating services require provenance information to not only attribute content but also offer additional services such as ranking of popular blog posts.

Technical Challenges:

Enable Trace back and correct attribution without violating user privacy and privacy laws
Disambiguating content authors with incomplete provenance information
Extend existing vocabulary for representing posts, such as SIOC, to model finer granularity provenance information.

Existing Work

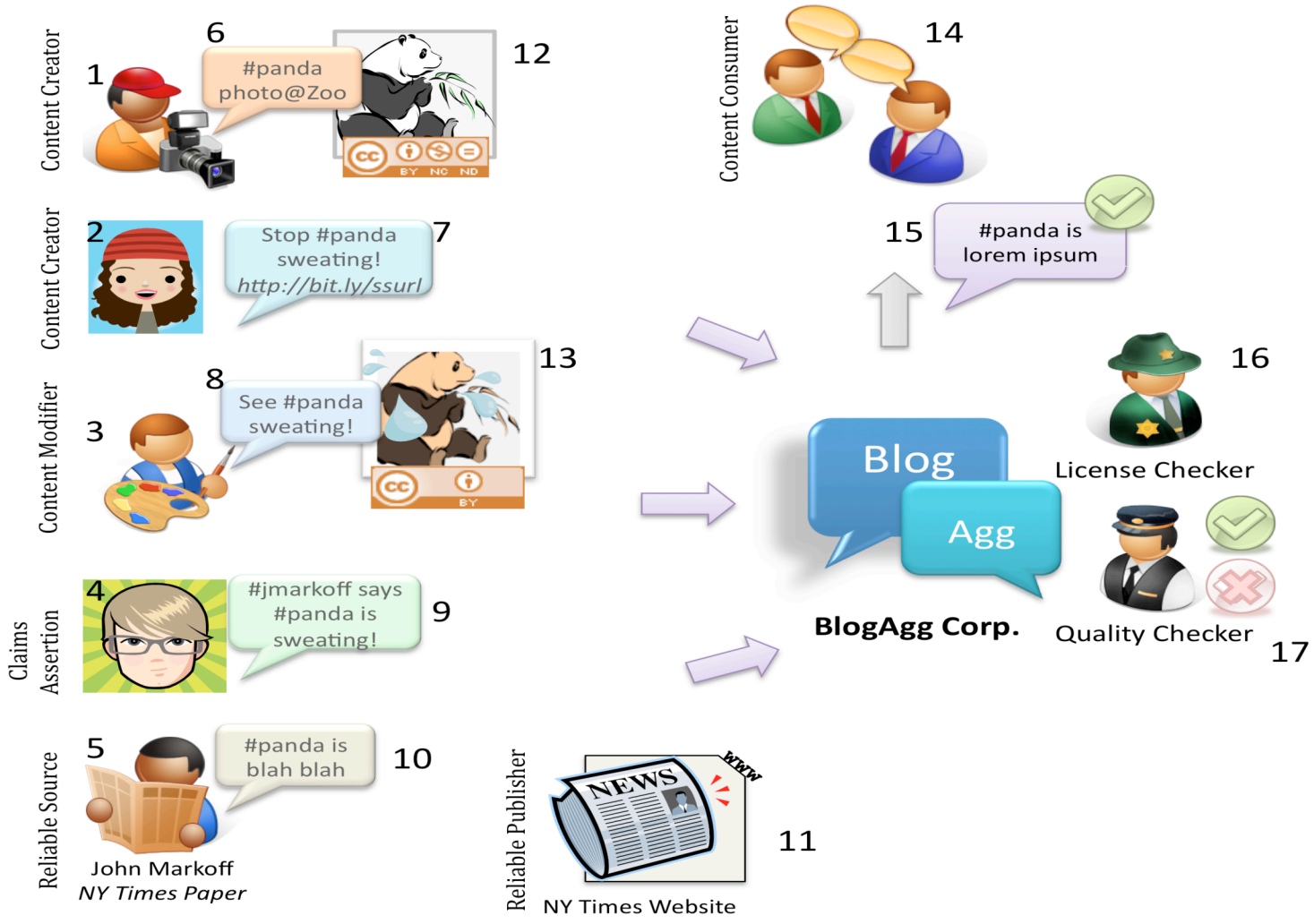
The SIOC project has developed a vocabulary for representing posts. This vocabulary is often used together with FOAF (that represent information about the physical person related to a sioc:User, e.g. its name, lastname, phone, social network, etc.) and SKOS, used mainly to represent topics and taxonomy relationships between these topics.

Three Flagship Scenarios for Provenance

1. **News Aggregator**
   - Theme: blog aggregator wishes to check the veracity of information published by other sources
   - Focus: web information, attribution, licensing
2. **Disease Outbreak**
   - Theme: government sources, scientific sources, and open web data are integrated and analyzed to create policies
   - Focus: e-Government, e-Science, linked open data
   - Issues: provenance for heterogeneous data integration, reproducibility through provenance
3. **Business Contracts**
   - Theme: Customer wishes to check that business delivered according to pre-defined contract
   - Focus: e-Business
   - Issues: provenance as proof, partial release of provenance records to respect IP

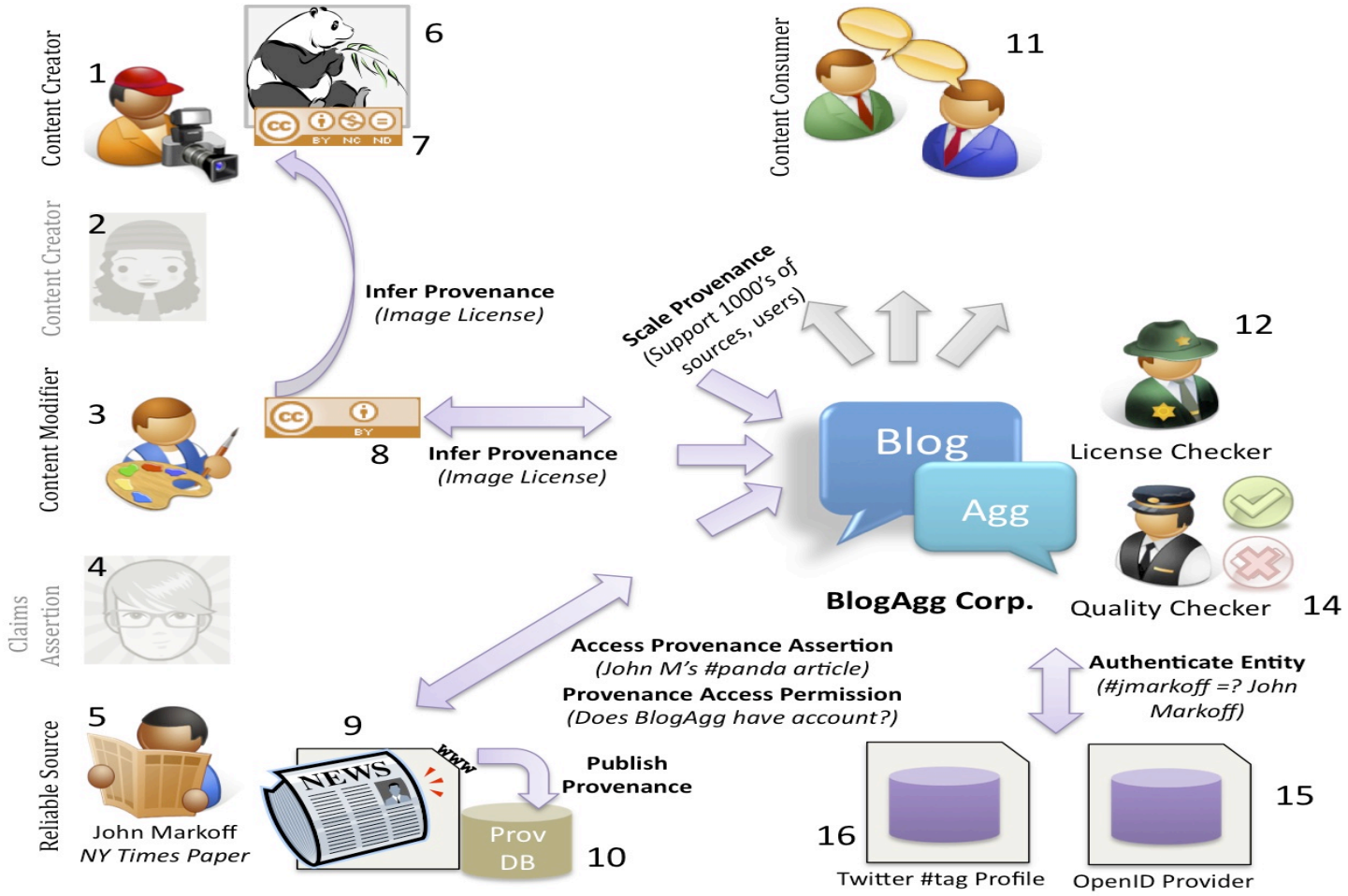
1) News Aggregator Scenario

BLOGAGG : CONTENT



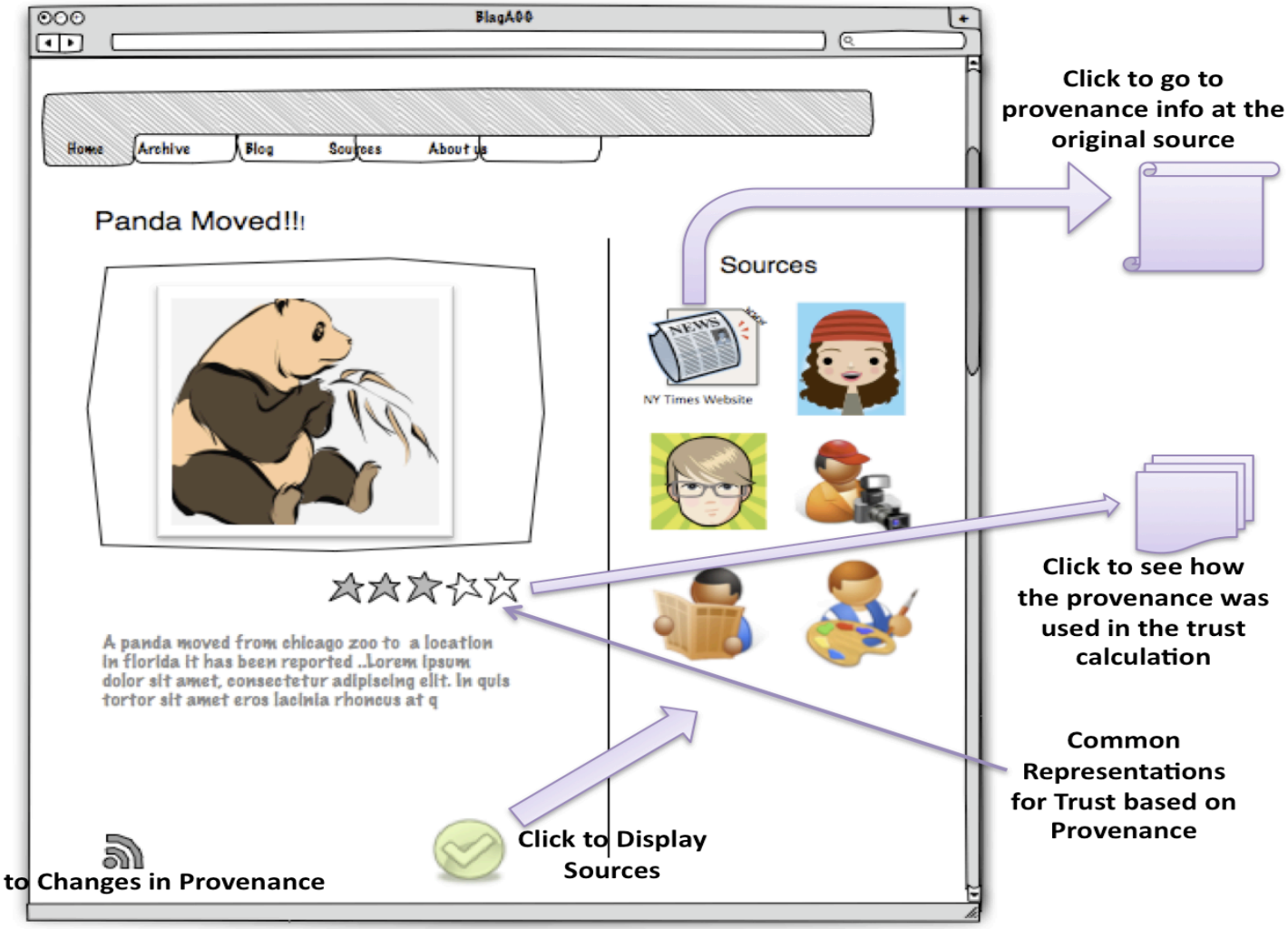
News Aggregator Scenario: Provenance Management

BLOGAGG : MANAGEMENT



News Aggregator Scenario: Provenance Use

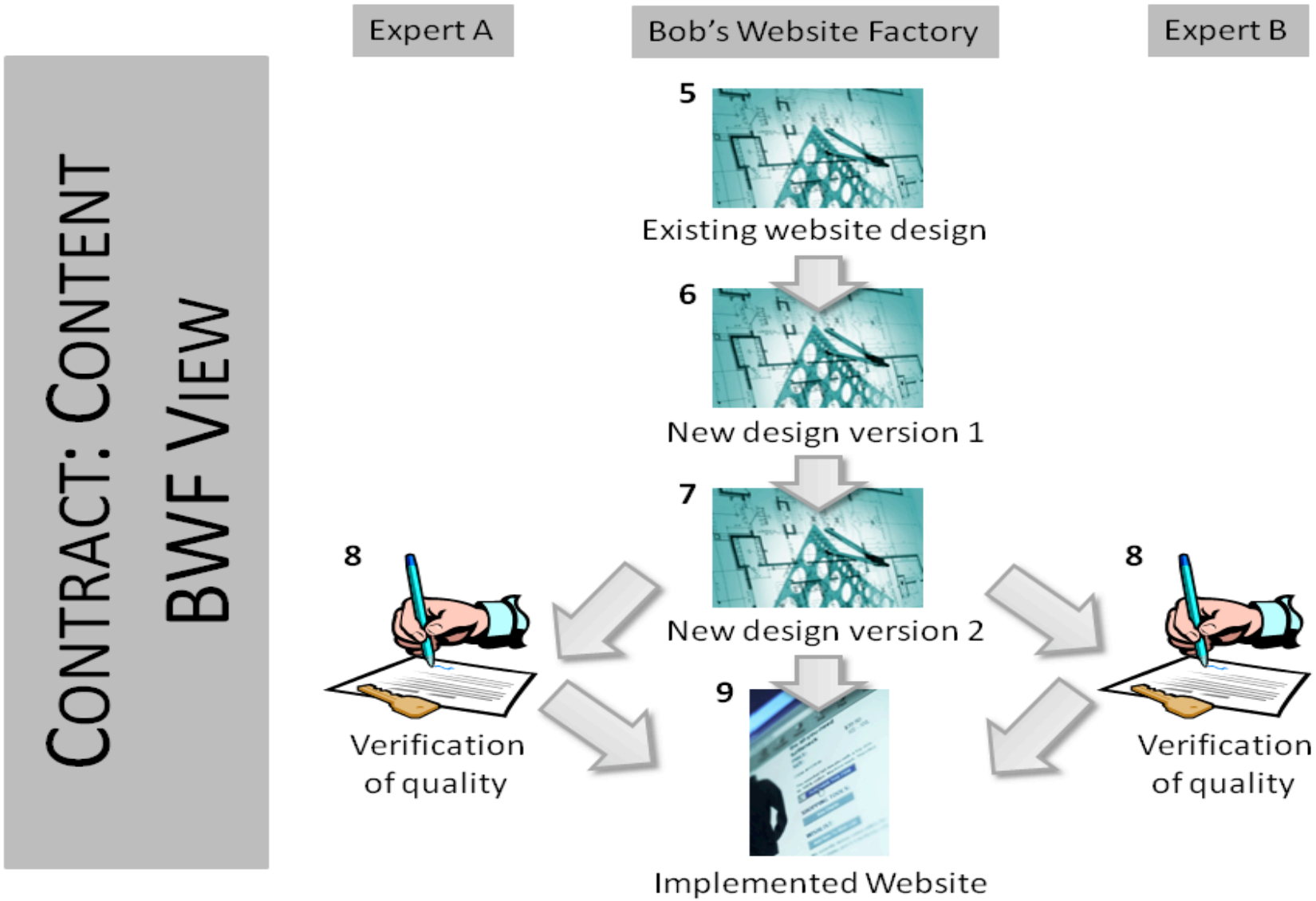
BLOGAGG : USE



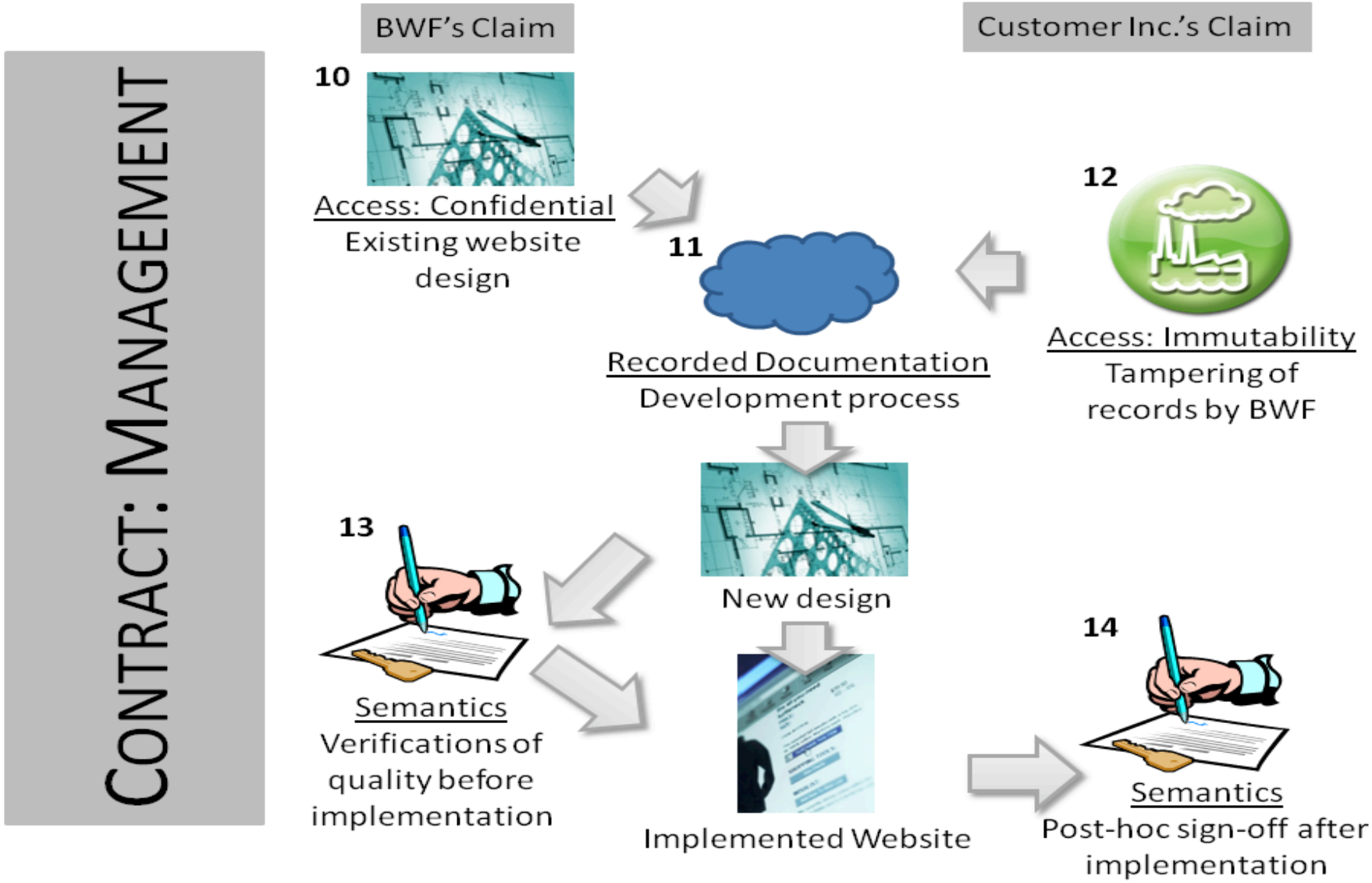
2) Disease Outbreak Scenario

- Data released by various sources (government, NGOs, news, blogs, etc)
- Social scientists integrate and analyze data, release new data about the spread of the disease
- Biologists analyze data as well, reuse data from social scientists, generate new results
- Analytical results are used by government to define policies to manage the dissemination of the disease
- Processes need to be repeated as new data becomes available over time

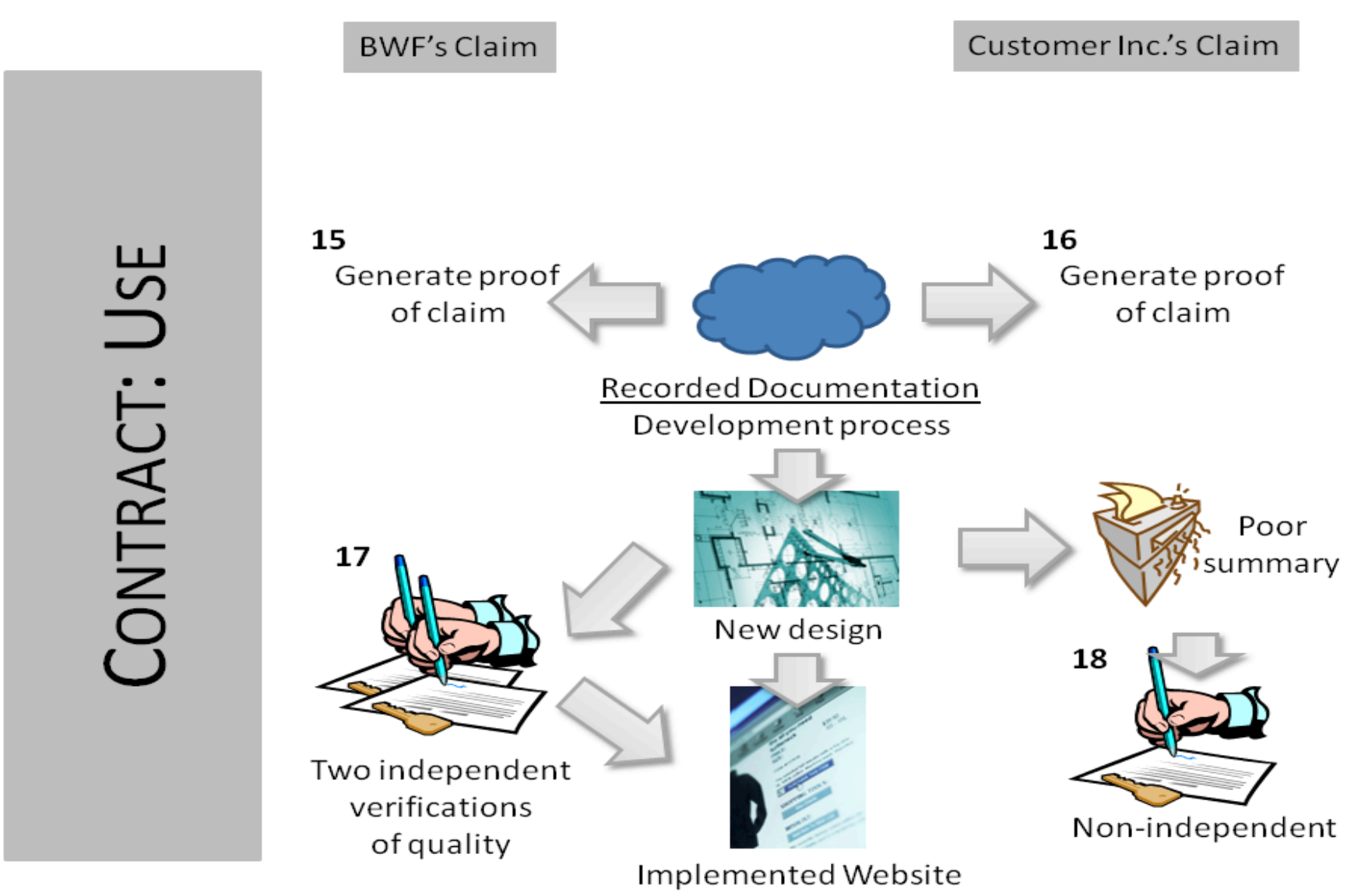
3) Business Contract Scenario



Business Contract Scenario: Provenance Management



Business Contract Scenario: Provenance Use



Outline

- Need for provenance on the web
- W3C Provenance Group: What is known about provenance in the community
  - Definition of provenance
  - Key dimensions of provenance
  - Use cases: Requirements for provenance
  - State-of-the-art and existing provenance vocabularies
  - Situating provenance on the Web architecture
- Towards a standard for provenance

State of the Art Report

- Created tagged bibliography collection
- Compilation of survey articles on provenance
- Invited presentations from people outside the group and in various communities
- Mapped 10 well-known provenance vocabularies
- Analysis of 3 flagship scenarios regarding state of the art
- Created gap analysis on what is missing to address the three flagship scenarios

Existing Provenance Vocabularies

- Many provenance vocabularies exist
  - Originated in scientific data management, workflows, library sciences, semantic web, etc.
- Provenance group mapped ten well-known vocabularies:
  1. Open Provenance Model (OPM)
  2. Dublin Core
  3. PREMIS
  4. WOT Schema
  5. Provenance Vocabulary
  6. Provenir ontology
  7. Proof Markup Language
  8. SWAN Provenance Ontology
  9. Semantic Web Publishing Vocabulary
  10. Changeset Vocabulary
- Used OPM as a reference model for the mappings
- Used SKOS when appropriate to express mappings

Provenance Spans Many Areas of Research

- Web
  - Tracking content dissemination on the web
- Workflow Systems
  - Computations leading to new data products, reproducibility
  - Reproducibility
- Databases
  - Query derivation, aggregations of data, streaming
- Knowledge representation and reasoning
  - Justification and explanation of reasoning
- Argumentation
  - What is taken into account to make a judgment
- Information retrieval
  - Question answering when documents are contradictory

Outline

- Need for provenance on the web
- W3C Provenance Group: What is known about provenance in the community
  - Definition of provenance
  - Key dimensions of provenance
  - Use cases: Requirements for provenance
  - State-of-the-art and existing provenance vocabularies
  - Situating provenance on the Web architecture
- Towards a standard for provenance

Provenance and Web Architecture: Provenance Situations to Consider

1. SIMPLE OH YEAH SITUATION
   - User retrieves a document, then clicks on "oh yeah" button, then site returns a provenance record
2. LICENSING SITUATION
   - User retrieves a document (eg an image), then wants to check permission to use
3. REFERRAL SITUATION
   - Site refers queries about provenance in terms of pointers to other site's provenance facilities
4. REPEATED QUERIES SITUATION
   - Service repeatedly queries a site, wants provenance for all the answers
5. VERSIONING SITUATION
   - User retrieves a document, then wants to see its provenance, but the document has been updated in the original site (its provenance as well)
6. DYNAMIC RESOURCE SITUATION
   - User retrieves a resource that is dynamically created

Outline

- Need for provenance on the web
- W3C Provenance Group: What the community understands about provenance
  - Definition of provenance
  - Key dimensions of provenance
  - Use cases: Requirements for provenance
  - State-of-the-art and existing provenance vocabularies
  - Situating provenance on the Web architecture
- Towards a standard for provenance

Towards a Standard for Provenance

- Currently discussing the charter of a Provenance Working Group
- Agreement on the following objectives:
  - Define a provenance exchange language and protocol to publish and access provenance
  - The scope of this language will be any resource, not be just semantic web objects
  - The exchange language should have a low entry point to facilitate widespread adoption, therefore it should be easy to do simple things
  - It should have a small core model and allow for extensions (ie, species/profiles, integration of other more expressive/complementary vocabularies/frameworks)
  - Some deliverables should be released early, WG should end in 18 months or 2 years

THANK YOU

W3C Provenance Incubator Group

Semantic Web Activity
World Wide Web Consortium

http://www.w3.org/2005/Incubator/prov/wiki

Special thanks to contributing group members: Chris Bizer, James Cheney, Sam Coppens, Kai Eckert, Andre Freitas, Irini Fundulaki, Daniel Garijo, Yolanda Gil, Jose Manuel Gomez Perez, Paul Groth, Olaf Hartig, Deborah McGuinness, Simon Miles, Paolo Missier, Luc Moreau, James Myers, Michael Panzer, Paulo Pinheiro da Silva, Christine Runnegar, Satya Sahoo, Yogesh Simmhan, Raphaël Troncy, Jun Zhao