July 2020

Impact Evaluation of PROV

A Provenance Standard Published by the World Wide Web Consortium

Lyndsay McAteer
IMPACT SCIENCE, ON BEHALF OF KING'S COLLEGE LONDON & NEWCASTLE UNIVERSITY

EXECUTIVE SUMMARY

This impact evaluation was conducted by Impact Science on behalf of King's College London and Newcastle University. Expert consultation was conducted with personnel from three organisations; AstraZeneca (a leading global biopharmaceutical company), NASA/USGCRP (the Global Change Research Program, responsible for the US National Climate Assessment), and The Gazette (the UK official journal of record, providing information on statutory notices). Through these activities, the project was able to collect and collate important evidence to assess what kind of impact PROV has been having. The major impacts of PROV were remarked upon in detail. Impacts were then applied to the Research Excellence Framework impact classifications (https://www.ref.ac.uk/publications/panel-criteria-and-working-methods-201902/, hereafter REF 2019/02) and this was followed by an analysis of how they matched the criteria. The most significant impacts of PROV are elucidated below.

1. PROV IS A COMMON LANGUAGE THAT CREATES A STANDARD

In its essence, PROV follows a formula that traces the truth of provenance for each article or component. A standard is set. The digitised nature of the process means that errors and omissions are easy to spot and correct.

REF 2019/02 – Impacts on the health and wellbeing of people and on animal welfare. Applications of PROV at AstraZeneca have assisted in the development of safe drugs and pharmaceuticals. PROV ensures that standards are as high as possible, thenceforth enabling this.

2. PROV IMPROVES ACCESS TO INFORMATION THROUGH THE USE OF LINKED KNOWLEDGE AND THE KNOWLEDGE GRAPH

PROV has played a central role in enabling access to information and linked data. Clear data with an accessible audit trail can be easily found online.

REF 2019/02 – Impacts on public policy, law, and services. PROV data is an integral component of websites at The Gazette and the US Global Change Research Program (GCRP); https://www.thegazette.co.uk/ and https://www.globalchange.gov/

3. PROV WORKS IN THE BACKGROUND TO PROVIDE A CLEAR, ETHICAL, AND TRANSPARENT DATA SOURCE

Using PROV means that information can be verified and traced through the data record. This gives the data credibility and authenticity.

REF 2019/02 – Impacts on practitioners and delivery of professional services; enhanced performance or ethical practice. At AstraZeneca, USGCRP/NASA, and The Gazette, PROV has led to improved practices and exceptional service. In terms of planetary science, the use of PROV has led to an increase in detailed, verifiable information on the Voyager project and the exoplanets programme.

4. PROV ENSURES THAT INFORMATION RELEASED TO THE PUBLIC DOMAIN IS ACCURATE

Accuracy is ensured through the continued use of PROV. Every segment of information can be broken down to its constituent parts, which means that an audit trail is readily available.

REF 2019/02 – Impacts on the environment. ***The National Climate Assessment reports of 2014 and 2018, released by the USGCRP, were completely reliant on the accuracy of the information provided. They needed to be able to withstand the scrutiny of a powerful industry lobby that sought to question climate science.***

Table of Contents

1. INTRODUCTION 4

2. IMPACT OF PROV 5

2.1 Accessibility (and ease of use) 5

2.2 Machine-Readable Data 5

2.3 Transparency 6

2.4 Accuracy 6

2.5 Credibility, Trust, Integrity 7

2.6 Reliability 7

2.7 Time Saving Efficiencies 8

2.8 Quality of the Data 8

2.9 PROV: An extensible platform 9

2.10 Potential for the Future 10

3. IMPACT CASE STUDY: THE GAZETTE 11

4. IMPACT CASE STUDY: ASTRAZENECA 12

5. IMPACT CASE STUDY: NASA/USGCRP 13

6. IMPACTS WITH REFERENCE TO THE RESEARCH EXCELLENCE FRAMEWORK (REF 2019/20) 15

6.1 Impacts on the health and wellbeing of people and animal welfare 15

6.2 Impacts on commerce and the economy 15

6.3 Impacts on public policy, law, and services 15

6.4 Impacts on production 16

6.5 Impacts on practitioners and delivery of professional services, enhanced performance, or ethical practice 16

6.6 Impacts on the environment 16

6.7 Impacts on understanding, learning, and participation 16

7. CONCLUSION 17

8. APPENDIX: METHODOLOGY 18

8.1 AstraZeneca 18

8.2 The Gazette 19

8.3 NASA 20

1. INTRODUCTION

Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness. This information is typically obtained by tracing the relationship between agent, entity, and activity, and the time that the activity started and ended.

Academic research on provenance has led to the development of a computerised standard called the PROV data model, or PROV-DM (https://www.w3.org/TR/prov-dm/). The PROV standard is published by the World Wide Web Consortium and its ongoing development has been enabled by research contributions from staff at King's College London and Newcastle University.

PROV was first published as a set of recommended standards in 2013 (https://www.w3.org/2011/prov/wiki/Main_Page), which includes a data model and an XML schema; an OWL2 ontology that maps the model to RDF; and a system to map that ontology to Dublin Core. It also includes a standard that makes it easy for anyone to access and query the provenance.

Significant impact has been evidenced through the use of PROV. This evaluation will draw on qualitative information to describe the impact. Some of the evidence will be provided through secondary research findings. There will be an emphasis on the impact of PROV outside academia, with a focus on three organisations where primary research was conducted; Astra Zeneca; The Gazette; and NASA. The research team has spoken with people at all of these organizations.

Specific impacts will be considered in respect of the Research Excellence Framework. Assessments and conclusions will be made that reflect this.

2. IMPACT OF PROV

2.1 Accessibility (and ease of use)

Because PROV data is stored electronically and catalogued by a specific standard, its access is not limited in any way. Open and linked data allows multiple organisations to share provenance information.

Steve Hughes, Principal Computer Scientist at JPL/NASA refers to PROV as a common language that works across disciplines. Furthermore, Hughes summarises the impact of PROV:

"It improves communication, it improves interoperability and it improves the ability to do science; it is all about communication."

Steve Hughes—13.2.20

PROV ensures that the digital records at The Gazette are always kept up to date. The Gazette is completely free to use and its on-screen display of the provenance trail enhances accessibility. Company information can be particularly useful for risk management and solvency details. It is easy to check to see if businesses are in trouble and to make commercial decisions with accurate information.

At the USGCRP/NASA, PROV data is used on their website at www.globalchange.gov. This provides public access to important reports and documents about climate change. PROV is an integral part of the connecting web of data.

New data will be incorporated into existing information through PROV. A common language and semantic framework will enable this. PROV makes it easier to track every specific step in the process to identify sources, measurements, or other data.

2.2 Machine-Readable Data

The machine-readable nature of PROV is directly connected to its accessibility and increased use across a wide range of sectors. Machine-readable data can be queried. PROV provides information that can be tracked in a meaningful way. At AstraZeneca, the digitised nature of PROV in the growth of nanopublications is commented on by Tom Plasterer, Director of Bioinformatics, Data Science & AI:

"PROV gets embedded within nanopublications and then some of the ontologies that support nanopublications right now. It's one of the really useful ontologies, you know, a number of them are either started from the W3C or branch off that W3C."

Tom Plasterer—5.2.20

Breaking down information into a more finely grained form offers greater scope for coding and producing identifiers that make data computer readable. This, in turn, increases the accessibility of data through PROV.

2.3 Transparency

The use of PROV indicates that due process is being followed, which correspondingly evidences transparency. With PROV, essential information is trackable and traceable. An example of this is the provenance trail that is displayed on the website of The Gazette along with every official notice. PROV will not allow any information to be buried deep down in the records. All constituent parts of the information record are apparent.

2.4 Accuracy

PROV applications are most useful where a high degree of accuracy is essential. In the case of our three flagship developments, this relates to historical detail (The Gazette); pharmaceutical science (AstraZeneca); and geological, astronomical, and climatic science (NASA).

An example of when the accuracy associated with PROV had a substantial impact is related to the publication of the National Climate Assessment Reports in 2014 and 2018 by the USGCRP. The politics surrounding the authenticity of climate change led to significant pressure for absolute accuracy in terms of data. Curt Tilmes, Computer Scientist at the NASA/Goddard Space Flight Center, commented on how PROV was used:

"The process that we had developed with the Third National Climate Assessment was so strong that it enabled us to really incontrovertibly make a scientific consensus assessment of these things such that you really couldn't…there was nothing to sink your teeth in and say that you disagree with this."

Curt Tilmes—27.1.20

The importance of accuracy at The Gazette is underpinned by its status as the official record: The information contained within its platforms can be used as evidence in a court of law. PROV is an integral component of the information on The Gazette website. Tracing PROV, step by step, on the website provides a clear map as to the accuracy of the information.

There is strict protocol and governance about placing notices at The Gazette. Business and Operations Director Janine Eves describes the process on placing notices in more detail:

"It's imperative, from the publishing perspective, that both notices and the notice placing authorities, or those people who place the notices, have been notified and they are authorized to place that particular information on the UK official public record, and governance around the whole process is absolutely critical.

They need to have a notice published, but actually, perhaps, they have changed gender, or they have come out of a relationship, so there is a problem with that information being published. We redact certain things and you can actually see the steps that we did in the provenance to redact."

Janine Eves—30.1.20

2.5 Credibility, Trust, Integrity

The very existence of the provenance information gives credibility to the data. It means that the data has integrity and legitimacy. When NASA constructs a satellite, it is made up of a large number of constituent parts. It is vital that the scientists and engineers know the history behind each part. Counterfeit parts pose a particular problem for NASA and it's a problem that is increasing. Between 2005 and 2008, according to US Customs notifications, the number of incidents of counterfeit parts increased from one to 604. The application of PROV could have a significant impact in terms of establishing the integrity of components, according to Curt Tilmes, who highlights the issue of parts coming to NASA via a long chain of intermediary suppliers and the utility of PROV in working back through that chain to identify the source of counterfeit or faulty parts.

PROV creates trust. At The Gazette, the use of PROV is a contractual requirement. It is accepted as something that is needed for reasons of trust. There is now greater trust in the data due to its use. People accessing The Gazette are often not interested in how PROV works. What concerns them is whether they can trust what is in The Gazette. Because The Gazette is the official record trust is a part of its wider definition.

PROV based data sources are trustworthy because they are verifiable. When an absolute level of trust is required, PROV can be relied upon. The confidence increases over time.

2.6 Reliability

Confidence, trust, and reliability are all interrelated. PROV has proven itself to be definitive and accurate across different fields. Because of its grounding in a real- world view of the world and its evidence base, it is gaining support and popularity.

"I think the point is that you get different levels of scepticism and I think in some cases the people always question whether the science is factual, and so from that sense, yes, provenance is clearly built to ensure that the evidence that has been used or brought forward to prove a point, to prove a theory; that the evidence that is reproducible, and if it is not reproducible then it raises the question more than once if that's where it fails, so I would say absolutely and so that does impact general policy, especially at NASA. This year NASA has got an increase in the budget of 12%. Now if we weren't doing good science then we would not be able to do that. Congress decided they would give us additional funding as a result."

Steve Hughes—13.2.20

2.7 Time Saving Efficiencies

Consulting PROV archives is quicker because a computer can search a whole database using a common language. Before PROV, tracing information could be a long and laborious process. Without PROV, it is unlikely that the 2018 National Climate Assessment Report by USGCRP would have been published to the same standard. In some respects, PROV has been pivotal in terms of its impact.

"That is all tough because my assumption is that we would not have actually done all if we did not have PROV. It is not that we wouldn't have done it, but we just would not have accomplished as much is probably the realistic answer."

Reid Sherman, Global Change Information System Team Lead—7.2.20

2.8 Quality of the Data

- Clearly, the quality of the data is respected at the highest level, as is evident from the use cases explored in this report.
- PROV is embedded within nanopublications, which are the smallest unit of publishable information that can be identified and attributed to its author (http://nanopub.org/wordpress/). The use of nanopublications, and its associated incentives and reward system, is encouraging information sharing within the scientific community. The use of PROV ensures that the data has trust, integrity, and credibility.
- In the Life Sciences and Data Infometrics sectors, the Allotrope Foundation is an industry wide consortium, founded in 2012, which is working towards improving communication and data compatibility. PROV is at the centre of their approach to scientific data: the Allotrope Data Format (http://docs.allotrope.org//TR/adf-audit/ADF%20Audit%20Trail.html) incorporates the W3C PROV Ontology (PROV-O), which expresses the PROV Data Model [PROV-DM] using the OWL2 Web Ontology Language (OWL2) (https://www.w3.org/TR/prov-o/).

"What they are trying to do with Allotrope is come up with common data models and common standards for different experimental measuring devices; so, things like mass spectrometers and high-performance liquid chromatography devices and things of that nature. At the core of it is the Allotrope Foundation Ontology and part of what they used is PROV."

Tom Plasterer—5.2.20

- PROV is often referred to as one part of the puzzle. It needs to work in conjunction with other components and is not perfect; however, as a system for digitally coding provenance information, its benefits and impacts are widespread.

- NASA was able to create a provenance schema for the Voyager ISS geometric calibrated images. Information was passed on to the exoplanet scientist to assist with their analysis of these images, as detailed in the NASA/USGCRP case study below. The Voyager spacecrafts are enabling in-depth learning about our solar system and this contribution to planetary science is immense.
- PROV has reduced the amount of debate in terms of what information needs to be collected. In terms of developing the archive, it means that instead of debating, agreeing, and disagreeing, more time can be spent on coming up with technical solutions. The wheel doesn't need to be reinvented each time.

2.9 PROV: An extensible platform

Specialists and practitioners are generally happy with PROV and how it impacts their work. Reid Sherman at USGCRP said that PROV was only limited by its capacity.

"I think what I would say is the use of PROV could be expanded to achieve better outcomes. I think there is more that we could do…a wider reach beyond USGCRP products, so, to have connections between more different scientific things, scientific concepts in our climate change science world."

Reid Sherman—7.2.20

Potential extensions to PROV were remarked upon by Stephen Cresswell at The Gazette. He thought that PROV should be more deeply embedded into the fabric of the workflow and that there should be a smooth process for doing this. Mention is also made about the use of domain specific terms, more sub-classes in the data, and a larger vocabulary.

PROV is referenced in various ontologies. For an updated list, see https://blogs.ncl.ac.uk/paolomissier/2020/06/04/w3c-prov-extensions/. A few of them are listed here:

- Agreements ontology https://promsns.org/def/agr/agr.html
- DBPedia DataId https://wiki.dbpedia.org/projects/dbpedia-dataid
- The Australian Dataset ontology https://data.gov.au/data/dataset/data-gov-au-dataset-ontology/resource/dc586d4f-a3a5-4e00-abb4-128277356bed
- The GDPR Provenance ontology https://openscience.adaptcentre.ie/ontologies/GDPRov/docs/ontology
- The MEX Performance ontology https://lov.linkeddata.es/dataset/lov/vocabs/mexperf
- The Onyx ontology for emotions expressed by user-generated content http://www.gsi.dit.upm.es/ontologies/onyx/
- The PAV lightweight ontology for tracking Provenance, Authoring and Versioning https://pav-ontology.github.io/pav/

- The ProvOne ontology for scientific workflow https://purl.dataone.org/provone-v1-dev
- The P-Plan ontology https://www.opmw.org/model/p-plan/
- The Stream Annotation ontology http://iot.ee.surrey.ac.uk/citypulse/ontologies/sao/sao
- The Org Vocabulary https://www.w3.org/TR/vocab-org/

2.10 Potential for the Future

Specific expanded vocabularies are yet to emerge for PROV. The importance of linked knowledge and the knowledge graph in terms of enabling PROV cannot be understated. Curt Tilmes talks about the timing of the impact of PROV:

"I personally believe that in the future the impact will be bigger, but I would definitely say that today we are not yet dependent on it. It is next generation that we are looking forward to seeing how we could use this.

But because they have encoded that data and it becomes kind of a chicken and egg thing, you can't really rely on it until everyone has their data on the knowledge graph or you are only going to find a small amount of things, but because we have kind of turned that point and we are starting to encode our data maybe the next people will encode their data in that way and next people will encode their data."

Curt Tilmes—27.1.20

Curt Tilmes explains that in the future we should be able to trace the provenance of all the food that we buy by scanning items with our phone. This would depend on all the information for every item being encoded. Perhaps this level of traceability would only be possible with advanced digitised technology.

PROV certainly has the potential to improve things for the scientific community in the future.

"There have been a lot of scandals and difficult thinking in the scientific community about reproducibility and about accuracy and integrity, and I think PROV is a structure that can really support advances in general scientific practice to address those concerns."

Reid Sherman—7.2.20

3. IMPACT CASE STUDY: THE GAZETTE

The Gazette has a contract with National Archives and is the source of official information in terms of statutory notices. It has been in existence since 1665 and has been the UK's official record since then. The Gazette also holds the contract to publish legislation (www.legislation.gov.uk).

The information presented in The Gazette carries legal weight. As a result, there is a necessity for an accurate provenance trail. The Gazette incorporates a diagrammatical representation of the PROV trail for each entry. Not everyone will want to see the PROV trail when checking a notice but the fact that it is on the website gives people full confidence in the information.

In 2019, The Gazette underwent a complete digital transformation. Prior to that it was published in both paper-based and web-based formats. The PDF or paper-based notice is now an exact copy of the online notice. Digitisation meant that PROV helped to define the online version of The Gazette. The provenance trail for each notice gives a clear representation of PROV. The number of users who visit the website and use this tool is increasing year on year.

Year	No. of users
2019	13,900
2018	8,495
2017	9,308
2016	10,595
2015	9,433
2014	6,958

The Gazette is an invaluable tool because of the degree of surety that it provides. From company research to tracing the war records of relatives, its information is considered reliable and authentic. The shift to a digital platform has enhanced its appeal and broadened its scope. The credibility, trust, and integrity of the data has been strengthened because of PROV. Its use is embedded within the data systems of National Archives, which results in numerous efficiencies due to machine reading and computerisation.

4. IMPACT CASE STUDY: ASTRAZENECA

AstraZeneca is a global pharmaceutical company with a major UK presence (its total global revenue is $24.4 billion, with 6500 employees in the UK) (https://www.astrazeneca.com/content/dam/az/Investor_Relations/annual-report-2019/pdf/AstraZeneca_AR_2019.pdf). PROV is now firmly embedded within the structures and systems of this organisation. It is seen as a key component and not usually considered in isolation. Its vital importance covers a wide range of processes from drug discovery, target identification, target validation, through to trial design, evaluation, clinical trials, and how data is managed in research and development. PROV ensures something akin to a gold standard in terms of the governance of data within this context.

It was first used at AstraZeneca to create dataset records and data catalogues for external data that was being imported into the company through their work on a project called CI360 (Competitive Intelligence 360). A knowledge graph and linked data were important components of this work. In pharmaceutical science, PROV provides automatic updates through the knowledge graph. This new information is then used to update internal assets around such applications as genomic studies, proteomic studies and metabolomics studies on clinical data. Being able to use PROV to capture this when the data came in and to establish who is responsible for them is invaluable.

The development of nanopublications has been a significant impact of PROV, with AstraZeneca having taken a central role in promoting this concept. Because these research and discovery publications are fully digitised and coded using the knowledge graph, scientific data becomes more available and accessible. Furthermore, AstraZeneca wraps each nanopublication in a small API so that it is embedded as a button within any application.

The development of the FAIR data movement (https://www.go-fair.org/fair-principles/) and the use of a semantic web technologies approach is something that has only come to fruition recently. In pharma science, some precursor systems existed that started generic annotation and the modelling of external clinical trial information, but they still used a linked data approach.

5. IMPACT CASE STUDY: NASA/USGCRP

NASA is the National Aeronautics and Space Administration. It is an independent agency of the United States Federal Government, and is involved in a diverse range of sectors including astronomy, planetary science, astrophysics, and earth science. PROV is operating at many levels throughout NASA, and it has been responsible for some significant scientific milestones and achievements.

USGCRP is a confederation of the research arms of 13 federal agencies, including NASA, which carry out research and development, and maintain the national response to global change in the US. At USGCRP, PROV is used in the Global Change Information website at www.globalchange.gov. This platform is open access and provides current data about the effects of global climate change. Metadata is structured, catalogued and organised with full provenance as each citation of each segment of information is tracked down to the author and the affiliations of the citation.

The National Climate Assessment (NCA) Report is published by GCIS and is available at the global change website. The third NCA report was published in 2014 using PROV. Every single finding that the report mentions gets its own identifier, and each identifier can link specifically back to the papers that support it. Each paper captures the data that was used within it, including the names of people and organisations.

The fourth NCA report was published in 2018, and it built upon the data collected for the 2014 report. The significance of PROV in this context is due to the political environment in which climate change research is conducted, particularly in the US. The second NCA had been published in 2009, and the contributing organisations were subject to a substantial number of vexatious Freedom of Information requests from political organisations opposed to mainstream climate science, who sought to cast doubt upon the credibility of the report. As a result, PROV was introduced for the third NCA report, to ensure that data provenance could be quickly and rigorously established. Thus, PROV based data was able to authenticate the research and findings from the scientists. The impact of PROV could be seen, according to interviewees involved in the creation of the reports, by the consequence that the 2014 and 2018 reports were not subject to a similar mass of FOI requests: the accuracy of PROV rendering many FOI requests pointless.

In terms of planetary science, provenance is an integral part of the metadata about exoplanets. NASA has an official archive, PDS, for all planetary science data, which is continually updated and verified through PROV (https://113qx216in8z1kdeyi404hgf-wpengine.netdna-ssl.com/wp-content/uploads/2019/05/130_crichton.pdf). PROV provides a common language and framework for knowledge, and has been incorporated into the PDS data model from 2019. This use of PROV extends to historical datasets from missions like the Voyager and Cassini, whose image data have been archived in this PROV-based system.

The Voyager programme, for example, was able to benefit from PROV because a provenance schema was created for geometric calibrated images. Images identified in the Planetary Data System archive were used for a curve analysis of one of the planets. This information was passed on to the exoplanet scientist who was able to use the derivations

with the images as a starting point for further analysis, as Steve Hughes explained when asked for an example of how he has used PROV in his work:

"I created a provenance schema for the Voyager ISS geometric calibrated images. So, these were images that we identified in the PDS archive and they are the ones that were going to be used for some type of curve analysis of the actual planet. So, what I did was to capture the provenance of how the images in the PDS were processed so that information can be given to the exoplanet scientist. They started the derivations using those images with analysis on those images."

Steve Hughes—13.2.20

Certain aspects of the work of NASA and USGCRP are gaining more and more public interest, planetary science and climate change in particular. This is reflected in the growth of media and social media articles. All this scientific information is verified and authenticated through PROV.

6. IMPACTS WITH REFERENCE TO THE RESEARCH EXCELLENCE FRAMEWORK (REF 2019/20)

6.1 Impacts on the health and wellbeing of people and animal welfare

- In 2016, the USGCRP used linked knowledge and PROV to produce a report that linked climate change with human health, 'The Impacts of Climate Change on Human Health in the United States: A Scientific Assessment' (available online at https://health2016.globalchange.gov/)
- Improved data governance and standards through the use of PROV has led to faster delivery of safer drugs for patient health and wellbeing. PROV means that companies like Astra Zeneca are able to get the most out of their data assets around drug discovery, target validation, trial design, and evaluation.
- ProvCaRe is one of the first publicly available repositories of provenance metadata extracted from published biomedical research studies, which can be queried through a web-based interface. PROV specifications allow the use of a large variety of semantic web technology tools. It is applied in clinical and healthcare research (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5977728/).

6.2 Impacts on commerce and the economy

- In 2020, NASA was given a 12% increase in its funding from the US Congress. This, in part, was due to its use of excellent scientific practices, which are grounded in PROV, although PROV is only one of many systems underpinning NASA's practices.
- In 2012, the Allotrope Foundation was launched as a pan-industrial organisation covering the biomedical, pharmaceutical, and biopharmaceutical industries. Key companies involved include GSK, Bayer, Biogen, and Pfizer. Allotrope is developing advanced data systems, which incorporate the linked data approach provided through PROV.
- Over the last ten years, nanopublications have emerged as an important source of scientific information. The data integrity that is provided through PROV is the key determinant of the growing popularity and success of nanopublications.

6.3 Impacts on public policy, law, and services

- At National Archives and The Gazette, PROV ensures that the official public record is credible and accurate. Information is trusted, and an audit trail is provided. PROV plays a significant role in updating legislation.

6.4 Impacts on production

- The use of PROV in the production of the 2014 National Climate Assessment Report meant that the subsequent report produced in 2018 was much quicker to complete because the previous knowledge was verifiable.
- There is anecdotal evidence that the use of PROV increases production. This is because the time efficiencies associated with its use, and the digitisation of data, which can be machine-read, will allow for processing of accurate information at a much faster speed.

6.5 Impacts on practitioners and delivery of professional services, enhanced performance, or ethical practice

- At The Gazette, the shift to web-based publishing led to the adoption of PROV as the standard ontology. PROV was adopted in order to deliver the level of credibility that is needed to carry legal weight for the Official Public Record.

6.6 Impacts on the environment

- The policy debate on climate change has been influenced through the work of USGCRP and Global Change Information System (GCIS). PROV informed research meant that all information was meticulously documented, which led to less controversy.
- The National Climate Assessment reports of 2014 and 2018 led to an increase of media and scientific citations. PROV was used to further understanding of climate change by using a solid and trusted process.

6.7 Impacts on understanding, learning, and participation

- The role of PROV in planetary science is a catalyst to the creation of new knowledge. Reliability and authenticity are guaranteed through the use of linked knowledge.
- At The Gazette, the Official Public Record notes the war record and commendations of service personnel. Recognising this historical record (informed via PROV) can assist with genealogical detail that is legally recognised.

7. CONCLUSION

Evaluating the impact of the PROV ontology at this time is like analysing a work in progress. This evaluation has described a range of outcomes and impacts that have resulted from the application of PROV. The W3C Working Group developed the PROV standard and produced the overview document from the W3C working group in 2013.

This impact evaluation has taken place over a two-month period up until the end of February 2020. The bulk of the findings have been extrapolated from interviews with seven individuals working within the three key organisations that we focused on. The resulting qualitative research due to this expert consultation was used to inform the research findings and to broaden the PROV knowledge base.

Desk research was conducted as part of this evaluation. Online articles and other pieces of information were consulted to expand knowledge of PROV. Some of this information was used when mapping findings against the Research Excellence Framework (REF).

A theory of change research methodology was adopted. Outputs, outcomes, and impacts were measured accordingly, with reference to REF 2019/02. This is detailed in the section, which looks at impacts within the REF framework. As far as was possible, research findings were applied, and examples were given.

A familiar thread from the interview respondents was that further extensions of PROV will be welcome. Interviewees expressed a need for more domains and a wider vocabulary for PROV. PROV has been successful in scientific and information management realms, but its wider application is where its future potential lies. This cannot happen until more data is on the knowledge graph. PROV is expressly designed to be extended in this way to specific domains and applications, as specific communities agree on the required extensions (https://www.w3.org/TR/prov-dm/#extensibility-section). Barriers to this happening are awareness of PROV and confidence in its outcomes/results.

Looking to the future, PROV has great potential. In order to fulfil this potential, it needs to have wider reach. Ongoing research and development are vital if PROV based applications are to gain more respect and greater coverage. This impact evaluation establishes that the PROV ontology is certainly meeting the criteria for impact in the Research Excellence Framework. This impact can be sustained and expanded with further research on and applications of PROV.

PROV provides the framework, but most importantly, it facilitates the communication of information. Steve Hughes from NASA (interviewed on 13.2.20) was correct when he said that it is all about communication.

8. APPENDIX: METHODOLOGY

A theory of change model was used in this impact evaluation. Stakeholder consultation was facilitated through a series of qualitative interviews with key personnel at each of the three flagship organisations.

Information about the impacts of PROV was collated and set out in the main part of the report. The Research Excellence Framework (REF 2019/02) was applied to these findings and an analysis of how the findings fit with the REF is provided. Theory of change will map impact with situation and circumstances throughout the report.

Interview tools were used to conduct skype and phone calls for the fieldwork part of the evaluation. The questionnaires follow.

8.1 AstraZeneca

PROV is a provenance standard published by World Wide Web Consortium that builds on research from Kings College London and Newcastle University. In this impact evaluation, we are looking for quantitative and qualitative descriptions of the benefits that have resulted from the application of PROV. The following questionnaire aims to provide more details about the benefits that have occurred due to PROV through your work at AstraZeneca.

I will be recording this interview to assist with the transcription. If at any time you don't wish to answer a particular question or if you want to stop the interview, just let me know.

Do you have any questions before we begin?

1. Can you tell me what your role (and job title) is at AstraZeneca? What kind of professional background do you have? [What size of organisation is AstraZeneca?]
2. How widely is PROV used within AstraZeneca?
3. In general, from what you have seen at AstraZeneca, what sort of impact do you think PROV has had?
4. Can you describe any specific impacts of PROV on the work that you do? Can you also explain how things worked before PROV was used?
5. What benefits have you seen using the PROV application?
6. If AstraZeneca is using PROV to make more information available through its website, is there any way that website hits (or usage) is being measured since it began to be used?
7. Can you measure or quantify any of the (other) tangible benefits that have occurred through PROV?
8. Outside of the scientific community, can you describe how the general public and key decision makers are influenced through the use of PROV?
9. Could the use of PROV be improved to achieve better outcomes?

10. Are there any elements of the PROV application that could be changed to achieve better outcomes for AstraZeneca?
11. Do you have any other comments that you'd like to make about PROV or its use, delivery, or impact?

8.2 The Gazette

PROV is a provenance standard published by World Wide Web Consortium that builds on research from Kings College London and Newcastle University. In this impact evaluation, we are looking for quantitative and qualitative descriptions of the benefits that have resulted from the application of PROV. The following questionnaire aims to provide more details about the benefits that have occurred due to PROV through your work at The Gazette.

I will be recording this interview to assist with the transcription. If at any time you don't wish to answer a particular question or if you want to stop the interview, just let me know.

Do you have any questions before we begin?

1. Can you tell me what your role (and job title) is at The Gazette? What kind of professional background do you have [What size of organisation is The Gazette (or National Audit office)]?
2. How widely is PROV used at The Gazette?
3. In general, from what you have seen at The Gazette, what sort of impact do you think PROV has had?
4. Can you describe any specific impacts of PROV on the work that you do? Can you also explain how things worked before PROV was used?
5. What benefits have you seen using the PROV application?
6. If The Gazette is using PROV to make more information available through its website, is there any way that website hits (or usage) is being measured since it began to be used?
7. Can you measure or quantify any of the (other) tangible benefits that have occurred through PROV?
8. Outside of the data science community, can you describe how the general public and key decision makers are influenced through the use of PROV?
9. Could the use of PROV be improved to achieve better outcomes?
10. Has the use of PROV at The Gazette led to a greater trust in the data? If so, is this measurable?
11. Are there any elements of the PROV application that could be changed to achieve better outcomes for The Gazette?

12. Do you have any other comments that you'd like to make about PROV or its use, delivery, or impact?

8.3 NASA

PROV is a provenance standard published by World Wide Web Consortium that builds on research from Kings College London and Newcastle University. In this impact evaluation, we are looking for quantitative and qualitative descriptions of the benefits that have resulted from the application of PROV. The following questionnaire aims to provide more details about the benefits that have occurred due to PROV through your work at NASA/USGCRP.

I will be recording this interview to assist with the transcription. If at any time you don't wish to answer a particular question or if you want to stop the interview, just let me know.

Do you have any questions before we begin?

1. Can you tell me what your role (and job title) is at NASA? What kind of professional background do you have? [What size of organisation is NASA?]
2. How widely is PROV used in NASA?
3. In general, from what you have seen at NASA, what sort of impact do you think PROV has had?
4. Can you describe any specific impacts of PROV on the work that you do? Can you also explain how things worked before PROV was used?
5. What benefits have you seen using the PROV application?
6. When compiling the 2018 National Climate Assessment report, information used in 2014 was updated accordingly using PROV. How much longer would it have taken to produce the 2018 report if this had not been used? Do you have any information of gained efficiencies?
7. Do you have any data on how PROV has reduced the amount of FOI requests relating to climate change?
8. If NASA is using PROV to make more information available through its website, is there any way that website hits (or usage) is being measured since it began to be used?
9. Can you tell me about the global change information system website – www.data.globalchange.gov and how PROV is used to keep this website up- to-date?
10. Can you measure or quantify any of the (other) tangible benefits that have occurred through PROV?
11. Outside of the scientific community, can you describe how the general public and key decision makers are influenced through the use of PROV?
12. Could the use of PROV be improved to achieve better outcomes?

13. Are there any elements of the PROV application that could be changed to achieve better outcomes for NASA?
14. Do you have any other comments that you'd like to make about PROV or its use, delivery, or impact?

RESEARCH OUTPUTS

- Desk research – articles and published information (online) about PROV
- Independent research – use of publicly available websites at The Gazette, https://www.thegazette.co.uk/, and the USGCRP Global Change website - https://www.globalchange.gov/
- Phone and Skype interviews with four people at NASA/USGCRP
- Phone and Skype interviews with two people at The Gazette
- Phone/Skype interview with one person at Astra Zeneca.

This case study was created by Impact Science, a Cactus Communications solution, on behalf of King's College London and Newcastle University.

www.impact.science

cactusglobal.com