CRIS in 2020

Keith G Jeffery
Science and Technology Facilities Council, UK

Summary

A vision for the requirements of a CRIS in 2020 is presented. It forms part of an e-infrastructure ICT environment for the whole domain of research. The history of CRIS is analysed to pick out key developments and the barriers to achieving the vision are addressed. A plan is proposed to meet the vision, overcoming the barriers and utilising expected developments in ICT. The key is CERIF used as metadata. The novelty is the concept of a CRISBOT.

1 Introduction

It is an honour and pleasure to be given the opportunity to address the CRIS community as outgoing euroCRIS president.

2020 is only 8 years away. The history of CRIS can be traced over 50 years, although CRIS used for management information, evaluation and presentation of research has only been realised in the last decade or so. Thus, 21 years after the first CRIS conference (1991) and 10 years after the first euroCRIS-organised CRIS conference (2002) we are at a point where assessment of our history and predictions of our future are appropriate. However, "prediction is very difficult, especially about the future" (Neils Bohr). He also produced another pearl of wisdom: "Never express yourself more clearly than you are able to think".

The paper is structured as follows: the vision is presented in section 2. Section 3 traces the history of CRIS and draws out key developmental stages. Section 4 considers the barriers to be overcome and section 5 concludes with a plan for euroCRIS to achieve the vision.

2 The Vision

2.1 Research

It is a human trait to be curious. Understanding the world around us – and indeed the universe – is a driving force and for millenia humans have addressed the process of understanding by research. Research consists of observation, experiment and modelling to create hypotheses and to substantiate them as theories. Research information and theories are communicated and subjected to peer review as a method of quality assurance. Over the years research has been aided by tech-

nology: telesopes and microscopes led to advanced particle accelerators and satellites. However the most significant technological instrument is the computer.

ICT (Information and Communication Technologies) provides the world of research with mechanisms for recording, organising, retrieving and re-purposing information. It provides mechanisms for accessing instruments and detectors and controlling them and the data they produce. It provides mechanisms for managing the processes (workflow) of research and mechanisms for intercommunication between researchers. It provides the ability to erect models of the world of interest and study their behaviour, comparing with observations and experiment. In short it extends greatly the power of the researcher or research manager, the innovator, the media and the citizen to accomplish previously unimaginable tasks. ICT is not a tool to assist the researcher like a telescope or microscope – it provides a novel way of doing research. This is well documented as 'The Fourth Paradigm' (Hey et al 2009).

2.2 The Vision

The vision concerns extending further this powerful research-supporting technology for the purposes of research and research management. From an end-user point of view the requirement is access from anywhere at any time using any appropriate device. The end-user should be able to interact in any mode (typing, speech, gesture, thought) as a dialogue with the ICT environment. The ICT environment should respond by assembling the relevant information, the relevant software services to do required processing and the required computing resources to execute any request within the dialogue. The requests may include the need for dynamic, real-time access to detectors and instrumentation to collect more data as well as utilising existing data. The requests may involve the creation of new software for novel tasks. The ICT environment should also – based on historical stored information and knowledge-based prediction – anticipate the end-user requests and interaction modalities. The presentation of the information to the end-user should be through appropriate modes, particularly output speech and graphics. CRIS – representing research management information – are an integral part of this ICT environment providing the context for the day-to-day work of the researcher – or research manager, innovator or the media.

The aim is virtualisation. The user neither knows nor cares from where the information comes nor how and where the processing is done as long as quality standards and service levels are maintained. The vision requires a user model, a processing model and a data model providing consistent services underpinned by a resource model. The key is metadata.

3 A Brief History of CRIS

3.1 Introduction

As an adjunct to a PhD in Geology, the author produced a software system (G-STAR) to manage research data. It was based on the relational algebra and included software for statistics, reporting, graphics as well as data management. It was used by fellow postgraduates and then – reengineered for production as G-EXEC – at the British Geological Survey (Jeffery & Gill 1974, 1976a) leading to ideas of an ICT environment for geoscientists (Jeffery & Gill 1976b). The user

interface was simple commands (EXEC, FIND, MAKE) with parameters – input on punched cards – which caused the composition of pre-written subroutines into a FORTRAN program. This dynamic composition and the relational data structure allowed for distributed parallel processing. It was used widely internationally in geoscience and then in wider environmental science and subsequently in some physical sciences. During this period G-EXEC was used as the 'switchboard' among multiple heterogeneous geoscience information systems utilising the Filematch Format (Sutterlin et al 1976). In 1980 the author was developing systems for research and was asked also to develop (from a pre-existing prototype) an online system for research grant management. At this point the author's career and CRIS intersect.

3.2 Pre-1980

The period before 1980 is characterised by batch processing systems of the funding organisations. These systems recorded awarded (sometimes proposed) research project grants usually with identifier, title, PI (principal investigator) and institution of the PI together with the amount awarded. Sub-records recorded the financial transactions.

3.3 1980s

In the 1980s there was a realisation of the value of research information and the potential of interoperability. The EC (European Commission) CREST Committee produced a report proposing interoperation (exchange) of research management data and a similar report was produced by the Conference of European Rectors. In parallel the IDEAS project (1984-1987 led by the author) (Jeffery et al 1989) demonstrated interoperation among research management information of UK, FR, IT and the subsequent EXIRPTS project (Naldi et al 1992) among the G7 countries plus Sweden. The EC convened a group of delegated national experts to develop a standard for interoperation of research management information and (CERIF91) was produced – albeit with objections from the author. CERIF91 had only the entity project and as attributes PI, institution, title. This meant that it could not represent accurately - for example - multiple PIs, multiple organisations or titles in multiple languages.

3.4 1990s

In 1991 Jostein Hauge from University of Bergen convened a group of experts in CRIS to the first CRIS conference. The group became known as euroCRIS. Subsequently throughout the 1990s the EC supported CRIS conferences in 1993, 1995, 1998, 1999 (jointly with the USA) and 2000. The author's papers in these conferences emphasise interoperation, data quality, validation, declarative query and reporting, use of WWW for the user interface and CRIS within e-Science. In parallel a group was convened to work on best practice for CRIS. Experience with CERIF91 demanded a revision. In 1997 the EC reconvened the expert group (with some personnel changes) and after much discussion a proposal for CERIF with formal syntax and inline semantics – developed by Anne Asserson and the author - was presented and accepted as (CERIF2000).

3.5 2000s

In 2002 the EC requested euroCRIS to take over maintenance, development and promotion of CERIF and euroCRIS was constituted formally. Conferences were held every two years starting in 2002. The author's papers (usually co-authored) concern workflows, relationship of CRIS to publication repositories, the place of CRIS in the e-infrastructure and utilising GRIDs with dynamic software services and also research evaluation. euroCRIS held member meetings and strategic seminars gathering strategic partners and influence. The task group structure ensured progress on all fronts. Notably the semantics of CERIF were taken as a separate layer and an XML variant for interoperation was defined. The developers of institutional repositories started to address the limitations of DC (Dublin Core) metadata with CERIF. The use of CRIS for evaluation and benchmarking increased alongside increasing use for management and decision support in universities. CERIF2000 and its subsequent releases became widely used and commercial offerings appeared.

3.6 2010s

Increasingly the definition of semantics became important. A strategic relationship with CASRAI was established. LOD (Linked open data) / semantic web emerged as a paradigm. The mapping from CERIF as a relational model to this paradigm was demonstrated, as was interconversion among several metadata formats using CERIF as the 'switchboard'. A LOD task group was created linked closely with the VOA3R EC project and in parallel VIVO (a US CRIS from Cornell University based on LOD) became a strategic partner. euroCRIS involvement in EC research and coordination projects led to significant developments and further promotion of CERIF. A second workshop on CRIS and repositories produced the (Rome declaration) which establishes joint evolution. By 2012 CERIF was accepted as the national 'standard' in 9 countries with many more organisations (funders, universities and research laboratories) using it. euroCRIS has members in all continents except Antarctica. CERIF is becoming accepted as a component of the e-infrastructure in EC-funded projects such as EPOS and as a component in the LOD/semantic web world for e-government open data through the ENGAGE project.

A particular development justifies recording. Following an agreement between universities, research funders, research administrators and researchers JISC (the UK organisation providing ICT services to the higher education sector) launched programmes in the area of research information management alongside their existing repositories programme and subsequent research data programme. This was justified by an independent technical report (JISC2009) which recommended CERIF subsequently backed up by a business case (JISC2010). This programme has funded multiple projects using CERIF to connect within and between universities and research funding organisations and has pushed forward the development of CERIF-CRIS considerably in UK and within the euroCRIS community.

The ambition for the 2010s must be to see CERIF-CRIS used universally, to see B2B (business to business) communication between CERIF-CRIS (especially between funders and research organisations), to have researchers and research managers using CERIF-CRIS as part of everyday life with easy-to-use interfaces, to see CERIF-CRIS used widely as the basis of benchmarking and evaluation and to see CERIF-CRIS as the gateway to all toll-free open access research information. Ideally CERIF will also be used to provide the gateway to a complete ICT environment for research.

4 Barriers

4.1 Introduction

The preceding section has charted very briefly the success of CRIS in becoming established widely and the success of CERIF as an interoperation standard. However, much remains to be done and there are significant barriers to achieving the vision described in Section 2.

4.2 Technical Barriers

The major technical barriers are ease of use, access, completeness, quality and competing architectures.

Present day CRIS systems usually have interfaces demanding keyboard and screen, and the vast majority are not compatible with mobile devices with their small screens and use of gestures for interaction. They do not exhibit intelligent user interfaces which adapt to a user profile and learn from user interactions. They do not use the semantic layer to propose improved requests e.g. more precise or broader queries. They do not anticipate user requests nor utilise the history of requests to assist the user interface.

Input of data can be improved greatly. Workflow should eliminate multiple requests for the same information by capturing it at first input by any actor within the system. Wherever possible automated metadata extraction from other information should be used to pre-complete input forms leaving the end-user to validate and add additional information uniquely owned by that end-user.

In general an end-user wishes access to their local CERIF-CRIS (and the research environment it gateways) and access to the world of CERIF-CRIS via their local system. This allows the same user model to be employed and provides a consistent user experience. Access to other CERIF-CRIS involves availability - is the other CERIF-CRIS known to the local CERIF-CRIS - and reachability - is there an appropriate (bandwidth, reliability) network path and appropriate transforming and translation facilities.

Completeness of information provision can only be achieved by ensuring the CERIF-CRIS ingests information from other systems (to avoid double input) and where possible metadata is generated from pre-existing information thus reducing the threshold barrier encountered by the end-user when inputting data. Quality can only be achieved by strong validation of input and update transactions and by display of relevant information for easy user validation. Techniques of comparison and statistical anomaly detection can be used.

Solutions for research management other than CERIF-CRIS have been developed. Progressively repository architectures are merging and co-evolving with CERIF-CRIS. CERIF-CRIS interoperate with systems based on LOD/semantic web. Similarly CERIF-CRIS generate web pages more effectively and efficiently than researcher authoring. CERIF-CRIS can be used to ‘front-end’ ERP systems for finance, human resources and other normal business functions.

4.3 Organisational Barriers

Most organisations have legacy systems that persist because it is too complex / expensive to replace them. Utilising a CERIF-CRIS to ‘front-end’ and integrate such systems then allows their

replacement without affecting the end-user perception of the total environment. This involves wrapping the legacy systems to 'talk CERIF' within the organisation but then provides a consistent architecture with all the intercommunication utilising formal syntax and declared semantics which provides flexibility and stability. The cost-benefit of utilising a CERIF-CRIS has been demonstrated by a JISC business case study (JISC2010) and increasingly is being seen in practice as more CERIF-CRIS are deployed. The lack of mandates in organisations - for employees to follow policies of workflow, open access etc relating to the CRIS – means that take-up is delayed but, as the benefits are better validated, mandates will appear more widely and there will be much greater employee compliance leading to better completeness and quality of the information in the CERIF-CRIS.

4.4 Human Barriers

The upside of human behaviour concerns intelligence and its use. The downside concerns stubbornness and strongly-held views. Overcoming this barrier requires evidence of benefit which appeals to the human intelligence and demonstration of competitive advantage by those who embrace CERIF-CRIS. Associated with the downside is the threshold effort to engage with the CERIF-CRIS. It is necessary to lower this barrier by intuitive, intelligent and helpful user interfaces with associated help / learning facilities to assist the end-user to overcome this barrier.

4.5 Legalistic Barriers

Security is an evergreen problem in ICT. In the case of CERIF-CRIS - which should be toll-free open access based on the perceived benefits of sharing information - there is less of a problem although data corruption by malicious users must be prohibited. The real problem concerns those parts of the CERIF-CRIS that are not toll-free open access, for example information on research related to national defence or experimentation on animals where the public interest is not served by availability. Similarly if the research leads to commercial exploitation the information - especially concerning patent applications – requires secure protection. Clear principles agreed internationally are needed.

A related concern is privacy. Medical and social science research commonly concern human subjects and thus personal data. Furthermore, unauthorised access could result in actions to the detriment of the individual or society. Here the balance between data protection (privacy) and freedom of information (transparency) needs to be managed and the 'acid test' is public interest for which clear principles need to be established internationally.

Not surprisingly national laws are not aligned on these issues giving problems in international access. Commonly it is illegal to process data concerning a citizen of nation A on systems in nation B. The US Patriot Act and the French law on digital information traversing its territory are examples of complications to be addressed. Without international agreement a global e-infrastructure will never be achieved fully.

5 The Plan

5.1 Breaking the Barriers

We need to address the barriers by subsuming them into a new environment for research and research management. This new environment proposed can overcome the technical, organisational, human and legalistic barriers. It also addresses the 'data deluge', the need for digital preservation and the new opportunities offered by the rapid development of ICT.

The technical barriers can be overcome by utilising (further developed) CERIF as the canonical syntax and semantics. This provides interoperation viewed from the 'home CERIF-CRIS' of an end-user. Furthermore assisting intelligent user interfaces with CERIF reduces the barriers. An overall architecture moving to virtualisation by autonomic computing allows declarative rather than procedural end-user requests to be satisfied. CERIF-CRIS implementations have already demonstrated integration with alternative architectures for research information such as repository architectures and LOD/semantic web and even with ERP systems.

The organisational barriers can be addressed by a superior architecture that wraps legacy systems allowing their timely replacement with 'best of breed' offerings without destroying the internal organisational interoperability. This improves the cost-benefit of adopting a CERIF-CRIS and this is further enhanced by utilising GRIDs for resource sharing and CLOUD computing which converts CAPEX to OPEX (capital expenditure to operational expenditure) in a 'pay as you go' manner already proved popular and cost-effective for mobile phone services. Full realisation of virtualisation requires dynamic (re-)composition of metadata described services to meet quality of service and service level agreements. Moving in such a direction encourages organisations to issue mandates to ensure policy is enacted by employees to the greater good of the organisation based on the experience of best practice.

The human barriers are addressed by providing benefits of direct relevance to the role of the end-user. Automated CVs, bibliographies, web pages and research proposals are a benefit to the researcher and the increased exposure leads to increased citations and potential collaborations. Customised management reports to funders and for internal strategic use benefit the research manger. The innovator sees appropriate opportunities for taking a research result through to wealth creation while the media have access to interesting research 'stories'. The general public – as well as obtaining information via the media – can become involved through 'citizen science' harnessing their enthusiasm to the greater benefit of all in activities ranging from searching for extraterrestrial life to searching for chemical molecules that could form the basis of a drug for common illnesses.

The legalistic barriers concern security, commercialisation, privacy and international law. All can be addressed with an appropriate ICT environment where metadata is the key. Holding the policies in these areas as metadata allows consistent application across information, software services, utilisation of computing resources and international aspects.

There are three key elements to achieving the required ICT environment: (1) the description of the ICT environment; (2) the interaction of the world of interest with any ICT system in that environment; (3) the interaction of the ICT systems with the user. All three rely on one key technology: metadata.

5.2 Architecture

5.2.1 Description of the ICT Environment

Concerning (1) the key is metadata which describes parsimoniously the complexity of the ICT environment of information, software and computing resources. Then it is possible to have autonomic computing where the end-user just specifies what is required (declarative request), not how to do it (procedural request). The implication is that the system environment locates relevant information and services, dynamically composes the services to meet the user request and locates appropriate computing resources to execute the request. Dynamic (re-)composition of services allows for distributed parallel execution thus achieving quality of service and service level criteria. Despite vendor lock-in in some ICT environments (public clouds are a prime example) an open market in software services with standardised metadata should allow interoperation above the infrastructure and platform levels so achieving virtualisation of the computing environment as seen by the end-user. Detailed work is necessary to assess if CERIF can describe adequately these software services as products and the computing environment as facilities and equipment together with the appropriate restrictions.

5.2.2 Interaction with the World of Interest

Concerning (2) the key is metadata which describes parsimoniously the complexity of the real world in a form suitable for the ICT system. The crucial aspects are formal syntax (to assure accurate ICT processing) and declared semantics (to allow logical ICT processing). The metadata needs to describe not only the information of relevance but also the associated restrictions on use (rights, privacy, security, any costs if not toll-free) of the information resources. A clean solution is for all information resources to be 'wrapped' by services as described above and for the services to take care of these aspects. CERIF acts as data for research management and simultaneously as metadata to more detailed information such as research publications, datasets and software, patents, products, facilities equipment, services, funding etc. It is here that CERIF used as metadata to describe the world of interest (information) intersects with a possible use of CERIF as metadata describing software (as products) or services.

5.2.3 Interaction with the User

Concerning (3) again the key is metadata. In this aspect it is necessary to overcome the organisational, human and legalistic barriers – integrated with the other two aspects above. In essence the end user wants to think about research, research management, innovation or communication of ideas and have the system assist automatically including anticipating the end-user's thought and proposing - in a dialogue - possible courses of action. Ideally this requires direct connection from the user's brain to the ICT environment, where the CERIF-CRIS forms the gateway.

This is probably not achievable by 2020 but should form the target for developments as steps towards this goal. From conventional interaction using keyboard, mouse and screen the first step - already seen with mobile devices – is to utilise gestures and voice. From simple gestures and voice commands the next step requires a user profile with history so the system can learn the user's intent (including standard operations) and react accordingly. An example is the move from the voice command 'call Mary' to 'I need a videoconference including shared desktops of the people involved in the project XXX' supplemented by 'I also need available to all participants

the current management information on project state including milestones, deliverables and resources'.

Similarly a researcher may wish to set up an experiment on a remote facility and needs to acquire access permission, timeslot, resources for controlling the experiment and data acquisition from her location and scheduling (and arranging travel and accommodation for) a subsequent visit for discussions with the facility scientists and technicians.

The key to all of this is the CERIF-CRIS acting as a gateway 'knowing' where all the resources are and managing the associated non-functional aspects (rights, security, privacy, performance). Already CERIF-CRIS are used to drive directories and webpages communicating research information and to provide the required research information. Turning this from a passive to an active function opens the way for more intelligent and responsive user interaction.

5.2.4 CRISBOT

From the above it is proposed that each user needs a software robot – a CRISBOT – to act on her behalf. The CRISBOT would manage all user interactions with the ICT environment and act as the intelligent assistant. The CRISBOT would mediate between the end-user current modes of interaction with the system and acquire progressively knowledge of the user habits and intent leading to helpful improvements of requests and providing proposals for requests that could or should be made by the user. The CRISBOT would manage the monitoring of progress of requests in the autonomic virtualised ICT environment, itself gatewayed by the CERIF-CRIS.

In time direct brain connections between the end-user and the CRISBOT would allow 'speed of thought' operations especially useful when the end-user is not only using the system for information processing requests (including modelling) but also to intercommunicate with others. At this point we would surely have achieved an ICT-based environment for research and achieving understanding – all based on CERIF-CRIS.

Acknowledgements

This contribution is the result of many years of working with extremely talented people in many fields of ICT and research whose contributions I acknowledge.

The euroCRIS Board and euroCRIS members represent the cream of that group in the CRIS domain.

In particular I acknowledge the work of Anne Asserson who had the courage (against the prevailing view) to co-develop the formal CERIF2000 model and who has continually criticised our work to try to improve. The result is seen in joint papers in previous CRIS conferences.

References

(CERIF91) ftp://ftp.cordis.europa.eu/pub/cerif/docs/cerif1991.htm

(CERIF2000) http://cordis.europa.eu/cerif/

(Dublin Core) http://dublincore.org/

(Hey et al 2009) The Fourth Paradigm: Data Intensive Scientific Discovery available freely at http://research.microsoft.com/en-us/collaboration/fourthparadigm/contents.aspx

(Jeffery & Gill 1974) K G Jeffery, E M Gill : 'G-EXEC, a Generalised FORTRAN System for Data Handling'. in 'Computer-Based Systems for Geological Field Data', Geological Survey of Canada Paper 74/663 - a State of the Art Report for 1973.

(Jeffery & Gill 1976a) K G Jeffery, E M Gill : 'The Design Philosophy of the G-EXEC System'. Computers and Geosciences 2(1976) 345-346

(Jeffery & Gill 1976b) K G Jeffery, E M Gill : 'The Geological Computer'. Computers and Geosciences 2(1976) 347-348 Paper presented at COGEODATA Conference, Paris, November 1975.

(Jeffery et al 1989) K G Jeffery, J O Lay, J-F Miquel, S Zardan, F Naldi, I Vannini-Parenti.: 'IDEAS: A System for International Data Exchange and Access for Science'. Information Processing and Management Volume 25 No 6 pp. 703-711, 1989.

(JISC2009) http://ie-repository.jisc.ac.uk/448/1/exri_final_v2.pdf

(JISC2010) http://www.jisc.ac.uk/publications/reports/2010/businesscasefinalreport.aspx

(Naldi et al 1992) Naldi F, Jeffery K G, Bordogna G, Lay J O, Vannini-Parenti I

A Distributed Architecture to Provide Uniform Access to Pre-Existing Independent, Heterogeneous Information Systems RAL Report 92-003

(Rome Declaration) http://www.eurocris.org/Documents/RomeDeclaration.pdf

(Sutterlin et al 1977) P G Sutterlin, K G Jeffery, E M Gill : 'Filematch: A Format for the Interchange of Computer-Based Files of Structured Data'. Computers and Geosciences 3(1977) 429-468.

Contact Information

Keith Jeffery

Director International Relations
Science and Technology Facilities Council

Rutherford Appleton Laboratory
Harwell Oxford
Didcot
OX11 0QX
United Kingdom

keith.jeffery@stfc.ac.uk