International Environmental Modelling and Software Society (iEMSs)
7th International Congress on Environmental Modelling and Software
San Diego, California, USA, D.P. Ames, N. Quinn (Eds.)
http://www.iemss.org/society/index.php/iemss-2014-proceedings

A Distributed Architecture for Sharing Ecological Data Sets with Access and Usage Control Guarantees

Ph.Bonnet [a]**. J.Gonzalez**[a]**, J. Granados**[a]

[a] *IT University of Copenhagen, Rued Langaard Vej 7, 2300 Copenhagen, Denmark*
(phbo@itu.dk,jgon@itu.dk,jogr@itu.dk)

Abstract: In our information-age, the necessary scientic exploration is mainly driven by mining new insights from many diverse data sets. While there is a consensus that a collaborative data infrastructure is needed to allow researchers in dierent domains to collaborate on the same data sets in order to get new insights, there are signicant barriers to the realization of this vision. One of the key challenge is to allow scientists to share their data widely while retaining some form of control over who accesses this data (access control) and more importantly how it is used (usage control). Access and usage control is necessary to enforce existing open data policies. We have proposed the vision of trusted cells: A decentralized infrastructure, based on secure hardware running on devices equipped with trusted execution environments at the edges of the Internet. We originally described the utilization of trusted cells for the management of personal data. We describe our vision and report on our progress towards the implementation of trusted cells on off-the-shelf hardware components. We show how trusted cells deployed in the field and throughout the community could make it possible to share ecological data sets with access and usage control guarantees. We rely on examples from terrestrial research and monitoring in the arctic in the context of the INTERACT project.

Keywords: arctic terrestrial research and monitoring, ecological data sets, trusted cells, data platform.

1 INTRODUCTION

The grand challenges facing our societies – climate change and energy supply – require significant scientific breakthroughs. In our information-age, the necessary scientitic exploration is mainly driven by mining new insights from vast, diverse data sets. The need to organize the management, publication and archival of these scientific data set, initially pointed out by Jim Gray Hey et al. [2009], is now widely recognized both in the European Union Wood [2010] and in the United States (most recently with the US President's Council of Advisors on Science and Technology pointing out the need for increased public access to federal biodiversity data Holder and Lander [2011])[1].

A collaborative data infrastructure is needed to allow researchers in different domains to collaborate on the same data sets in order to find new insights. While there is a consensus on the need for such an infrastructure, there are significant barriers to its realization. The final report of the High level Expert Group on Scientic Data submitted to the European Commission in October 2010 Wood [2010] summarizes the requirement for such an

[1] Note that while the problem of access and usage control is not directly mentionned in the PCAST report on ecosystem data, the committee has acknowledged the need for strong attribution of the data according to Alon Halevy from Google.

infrastructure as follows: *The emerging infrastructure for scientific data must be flexible but reliable, secure yet open, local and global, affordable yet high-performance*. There are thus many tensions to be resolved. In particular, the report mentions the problem of access and usage control: *the data must be available to whomever, whenever and wherever needed, yet still be protected if necessary by a range of constraints including by-attribution licenses, commercial license, time embargos, or institutional affiliation*. How do scientists share their data widely but also retain control over who uses them and how? This is the core problem we focus on.

In this paper, we restrict our attention on ecological data about arctic terrestrial ecosystems. While observation and monitoring data is expensive to obtain in the arctic because of the remote nature of the environment, the problem of access and usage control has never been particularly acute in this community. Indeed, research groups have had a form of monopoly about a given area. As a result, a data set about CO2 flux exchange originating from the Abisko station in Sweden, or from the Zackenberg station in Greenland could only have been obtained by a handful researchers. These researchers are administrating the data they collect and grant access to the data sets upon request. For example, the regulations for using Abisko Scientific Research Station (Abisko Naturvetenskapliga Station, ANS) climate data is a good representative of the state of the art in terms of data access in the arctic research community: *When you wish to use climate data collected under the responsibility of ANS please note that a formal request for data is needed. The request should be supplied by a brief outline of the intended use of the data. The data should not be passed to a third person* [2].

This situation is changing dramatically. First, the situation of monopoly in terms of data collection no longer exists. The INTERACT infrastructure project [3], funded by EU FP7, is pushing this paradigm shift. The Interact project regroups 33 arctic field sites. Its main goal is to promote transnational access to these sites, thus allowing scientists and policy makers to identify, understand, predict and respond to diverse environmental changes throughout the wide environmental and land-use envelopes of the Arctic. As a result, station managers are coordinating monitoring activities across sites and researchers are encouraged to conduct observations at several sites. The issue of attribution is thus no longer trivial. Second, offline access and usage control no longer fits the needs of the community in terms of data sharing. There is a growing pressure for scientists to share their data and make them available online. This was formalized in reports from the OECD on Principles and Guidelines for Access to Research Data from Public Funding in 2004 and 2007. These reports recommended to formulate explicit, formal institutional practices, such as the development of rules and regulations, regarding the responsibilities of the various parties involved in data-related activities. These practices should pertain to authorship, producer credits, ownership, dissemination, usage restrictions, financial arrangements, ethical rules, licensing terms, liability, and sustainable archiving. An instance of this open data policy is the Data Policy adopted for the International Polar Year (IPY) in 2006. It recommended that *Data should be accessible soon after collection, online wherever possible*. This is however not a reality yet. In their State of Polar Data Parsons et al. [2009], an assessment of the IPY data legacy, Parsons et al. remark that *overall, data sharing is commonly recognized as a scientific imperative, but the technical mechanisms require further development and cultural norms of science still resist sharing*. They add that *Data citation is increasingly recognized as a valid process, but implementation is sporadic at best*. Finally, Parsons et al. recognize the need for strong access and usage control mechanisms: *data should be as unrestricted as possible, but scientists need to establish norms of behavior that ensure proper, informed, and equitable data use. Good data policy helps move open data sharing forward, but it must be enforced*.

[2] `http://iasoa.org/iasoa/index.php?option=com_content&task=view&id=128&Itemid=140`
[3] `http://www.eu-interact.org/`

So, how to enforce access and usage control of data in the scientific community? An approach is to rely on guidelines and policies manually enforced by data managers. This approach was successful in the context of the US Long Term Ecological Research (LTER) Network Porter [2009]. The premise is that data managers are funded to enforce policies. This has not been the case in the polar community. This largely explains the lack of success of data sharing during the International Polar year, and the failure of the Polar Information Commons , supported by CODATA, and launched in 2009. Besides, it is virtually impossible for a data manager to verify that data is not transmitted to a third party, or that data is appropriately attributed in the litterature. Is it so possible to completely enforce access and usage control of scientific data, without the manual intervention of a data manager?

The hypothesis that underlies our work is that Privacy Enhancing Technologies can be used to automatically enforce usage and access control in the arctic ecologists community. Note that such mechanisms do not preclude the need for data managers; on the contrary, data managers could leverage these mechanisms and shift their focus from access and usage control to other stewardship tasks such as quality control or long-term archival. In this paper, we present our vision of data sharing based on a decentraliwed data platform that provides strong usage control guarantees and we report on the current status of our prototyping efforts.

2 VISION

Let us take a simplified example of how ecological data sets are collected and shared in the INTERACT community. Figure 1 shows activities taking place in a scenario involving two field sites (Zackenberg and Abisko), four labs (DMU, U.Lund, U.Oulu, and U.Alaska) and the online Knowledge Network for Biodiversity (KNB) repository for ecological data sets recommended by Nature. Data acquisition takes place at the field sites. Data acquisition is either manual or automatic; most of the collected data is digital on data loggers or PDAs, some of the data collected manually is in analog form (and is later digitized in a lab). On Figure 1, A and E are technicians collecting data in the context of monitoring programs at Zackenberg and Abisko, while B is a researcher from U.Oulu who is collecting observations during an expedition at Zackenberg.

Today, data is transfered from the field sites to a lab either via physical transfer of the storage media (sneakernet). A goal of the INTERACT project is to make the streams of data collected by data loggers available on the Internet. Note that data loggers cannot be directly connected to the Internet because of obvious connectivity limitations. Data is gathered via a low power local area network (LLN) onto a hub that is connected to a lab gateway via satellite connection (Iridium). The data collected from the data loggers are stored onto a database on the gateway, then made available online as a stream (views might be defined for the purpose of access control).

Let us now illustrate how data is processed and shared in the context of a monitoring program. M is manager of the monitoring program at DMU (NERI). He receives the data collected by A in the field. Her responsability is then to go through a data cleaning phase, and possibly a data derivation phase. This phases might be conducted together with colleagues from other institutions (U.Lund) in our example) participating in the program, under the responsability of the program manager. Today, data sets are exchanged by mail or made available on ftp sites, i.e., a weak form of access control is enforced and there is no usage control. Double checks are necessary to make sure that an anomly detection or a data derivation are conducted properly. This is very time consuming.

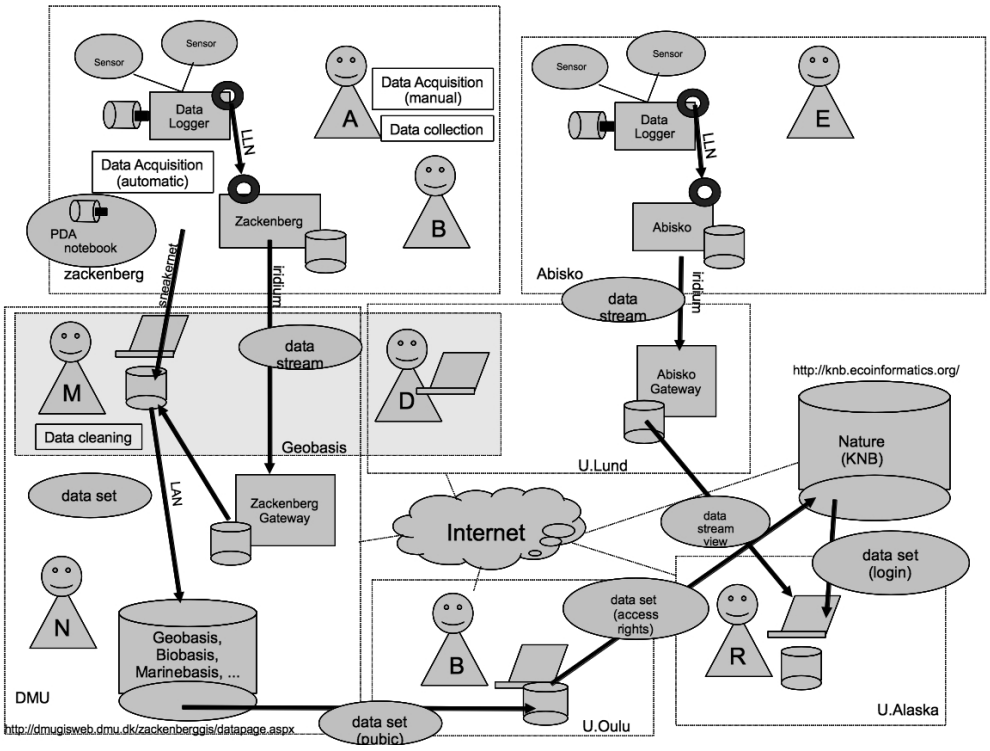

Figure 1. Sanitized case study illustrating data sharing in the INTERACT community

When the program manager is satisfied with the quality of a data set, she can publish it. To do so, she sends the data set to a local data manager who loads the data into a database equipped with a web query front-end. Once, data is made publically available on the DMU web site, the program manager has no longer control over who accesses it or how it is used. For example, if a new annotation is attached to a data set, there is no way to communicate it to the researchers who previously downloaded the data set. There is also no way to check that the researchers that use the data actually follow the usage policy defined by the data manager.

Let us now turn our attention to how data is processed and shared in the context of a research project. B collected data in the field. Back in the lab, at U.Oulu, he combines the data he has patiently gathered over the last 10 years with data sets he downloaded from the DMU web site. He builds a new predictive model and publishes his findings in Nature. As part of the Nature publication policy, he makes his data available online on the KBN web site. Ideally, reviewers and other researchers should be able to check that his model actually fits the data he gathered, they should be able to check that there is no obvious anomaly in the collected data, but they should not be able to derive a new model without explicitely crediting B for his work. Note that the access control mechanism provided by the KBN web site allows B to control who is accessing his data, but not how this data is then used (and possibly transmitted to a third party). For example, R from U.Alaska has contacted B to get access to his data set. Once access is granted, R cannot track how the data is used and how it is combined with the data stream obtained from Abisko. More interestingly, R might not credit DMU for collecting the data sets that were instrumental for building B's model. This is an illustration of the issue identified by the OECD report on open data policies: Whenever possible, access to data sets should be linked with access to the original research materials, and copied data sets should be linked with originals,

as this facilitates validation of the data and identification of errors within data sets.

Note that researchers might in fact be more interested in a form of auditing of how their data is used rather than in enforcing strict usage control policies, governmental organizations might share this goal, while private organizations might want to enforce more precise usage control policies. Our approach consists in providing a uniform solution to those different requirements. But, what is usage control? Usage control models usually refer to $UCON_{ABC}$ Park and Sandhu [2004]. In the $UCON_{ABC}$ model, subjects provide or consume data objects. A subject accesses objects via a set of usage functions referred to as rights. A reference monitor is responsible for taking usage decisions based on the subject and object attributes as well as Authorization, oBligations and Conditions (ABC). Authorizations are predicates that define whether a subject is authorized to hold a right; obligations are predicates that define the actions that a subject must take before or while it holds a rights; and conditions define predicates that must hold true about the environment in which the subject requires a given right. In our experience, current open data policies defined for ecological data sets can easily be expressed in terms of (i) rights, and (ii) authorizations, obligations and condition rules.

What does it take to enforce a $UCON_{ABC}$ model? The data platform should implement a reference monitor and guarantee that there will be no access to data objects unless appropriate usage decisions (i.e., decisions that respect the contract negociated by all parties) are taken. The basic idea is that the reference monitor is a software component that relies on hardware security features separating a secure world (where objects, attributes, as well as rights, authorizations and conditions are securely stored, while usage decisions are securely executed) and a non secure world (where the rest of the processing takes place). There is today, to the best of our knowledge, no implementation of any $UCON_{ABC}$ model.

How to enforce access control and usage control in this scenario? A solution is to consider that all actors rely on Google's FusionTable (or a similar centralized architecture) to store and share their data. The features of FusionTable might not exactly match the needs of all actors, but they come pretty close. A problem could be the lack of connectivity on the field sites, but this might not be a major barrier to adoption. A more serious problem would be the monopoly position gained by Google if all researchers and institutions started to publish their data on FusionTables. The solution we investigate is a different point in the design space. We define a distributed infrastructure for sharing ecological data sets with access and usage control guarantees. In this distributed architecture, researchers and data managers are equipped with trusted cells. We describe the architecture of trusted cells in the next Section.

3 TRUSTED CELLS

The decentralized architecture we propose for the sharing of ecological data sets with access and usage control guarantees is based on Trusted Cells interconnected via an Untrusted Infrastructure Anciaux et al. [2013].

Trusted Cells: A trusted cell implements a client-side reference monitor Park and Sandhu [2004] on top of secure hardware. At a minimum, the hardware must guarantee a clear separation between secure and non-secure software. We abstract a Trusted Cell as (1) a Trusted Execution Environment, (2) a tamper-resistant memory where cryptographic secrets are stored, (3) an optional and potentially untrusted mass storage and (4) communication facilities. Physically, a trusted cell can either be a stand-alone hardware device (e.g., a smart token) or be embedded in an existing device (e.g., a smartphone based on

ARMs TrustZone architecture).

The very high security provided by trusted cells comes from a combination of factors: (1) the obligation to physically be in contact with the device to attack it, (2) the tamper-resistance of (part of) its processing and storage units making hardware and side-channel attacks highly difficult, (3) the certification of the hardware and software platform, or the openness of the code, making software attacks (e.g., Trojan) also highly difficult, (4) the capacity to be auto-administered, contrary to high-end multi-user servers, avoiding insider (i.e., DBA) attacks, and (5) the impossibility even for the trusted cell owner to directly access the data stored locally or spy the local computing (she must authenticate and only gets data according to her privileges). In terms of functionality, a full-fledged trusted cell should be able to (1) acquire data and synchronize it with the users digital space, (2) extract metadata, index it and provide query facilities on it, (3) cryptographically protect data against confidentiality and integrity attacks, (4) enforce access and usage control rules, (5) make all access and usage actions accountable, (6) participate to computations distributed among trusted cells. Basic (e.g., sensor-based) trusted cells may implement a subset of this.

Untrusted infrastructure: The infrastructure provides the storage, computing and communication services, which expand the resources of a single trusted cell and form the glue between trusted cells. By definition, the infrastructure does not benefit from the hardware security of the trusted cell and is therefore considered untrusted. We consider that a Cloud-based service provider implements the untrusted infrastructure .

In terms of functionality, the untrusted infrastructure is assumed to: (1) ensure a highly available and resilient store for all data outsourced by trusted cells, (2) provide communication facilities among cells and (3) participate to distributed computations (e.g., store intermediate results), provided this participation can be guaranteed harmless by security checks implemented at the trusted cells side.

4 PROTOTYPE

Our first prototype, developed in the context of the INTERACT project, focuses on the storage and retrieval of time series in a trusted execution environment. This is the foundation of a trusted cell. The design of access and usage control policies adapted for the INTERACT community, and the design/implementation of mechanisms to enforce these properties is a topic for on-going work.

According to the GlobalPlatform consortium [4], a trusted execution environment (TEE) is a secure area of a computing device that ensures that sensitive data is only stored, processed and protected by authorized software. The secure area is separated by hardware from the device's main operating system and applications. Put differently, computing devices fall in two categories: (i) general purpose devices that do not provide any guarantee for authorized software, and (ii) secure devices, where authorized software can execute in a secure area which is not accessible by the rest of the system. Secure devices separate a secure area from a non-secure area, which we denote rich area in the rest of this document, following the terminology from the GlobalPlatform consortium, while general purpose computing devices merely provide a rich area.

A potential large number of applications are expected to be installed and run concurrently in the rich area, therefore sharing software and hardware resources (e.g., libraries, drivers, peripherals, memory, etc.). As in any modern rich execution environment, access

[4] http://www.globalplatform.org

control is in effect to protect shared resources, with a user space distinct from a kernel space. Software running in kernel space has access to all resources (e.g., data and programs in memory or on secondary storage). The rich area is in constant exposure to the outer world, thus attacks are to be expected. The assumption should be that with enough time and expertise an attacker can obtain root access to kernel space. As a consequence, all programs and data stored in or handled in the rich area are at risk of being exposed, including sensitive data and encryption keys. As we stated above, a secure device provides a hardware separation between the secure area and the rich area. Note that this is a slight generalization of GlobalPlatform's definition, which mentions that rich and secure areas should run on a same processor. We extend this definition so that a secure co-processor, or a smart card are considered trusted execution environments. In fact, we consider that any device equipped with a processor for the rich area, and another one for the secure area is a secure device as long as there is a guarantee that only authorized software runs in the secure area. We even consider that a virtual machine (VM) is a trusted execution environment as long as the hypervisor relies on hardware primitives to isolate each VM. Defining an ontology of these secure devices and precisely qualifying how secure they are is beyond the scope of this demonstration. It is a topic for future work.

Whenever rich and secure areas share physical memory, it is mandatory that software running in the rich area by no means is able to access memory allocated to the secure area. If peripherals are shared Secure must have prioritized access to them in such a way that if the rich area is compromised, the peripheral can still be accessed from the secure area even when a DoS attack is launched against it. More generally, the only assumption made in the secure area is that the authorized code that is run there can be trusted to protect the integrity and confidentiality of the data. This is a big assumption, but model-based development and formal methods can provide interesting guarantees. Again, exploring how model-based development can be leveraged in the context of secure data management is a fascinating topic for future work. Our prototype relies on ARM TrustZone. In [2] we presented a framework that combines commercially available hardware and open source software and enables the development of applications for TrustZone's secure area. In this framework, rich and secure areas are running on a single processor. TrustZone provides a hardware mechanism to separate the secure and rich areas. This mechanism relies on the so-called NS bit, an extension of the AMBA3 AXI Advanced Peripheral Bus (APS), a peripheral bus that is attached to the system bus using an AXI-to-APB-bridge. The NS bit distinguishes those instructions stemming from the secure area and those stemming from the rich area. Access to the NS bit is protected by a gatekeeper mechanism in the Operating System. The Operating system thus distinguishes between user space, kernel space and secure space. Only authorized software is running in secure space, without interference from user or kernel space. TrustZone defines a communication abstraction for the interaction between programs running in the rich area that act as clients, and programs running in the secure area that act as servers. This client-server communication is session-based (each session is bound to programs in the rich area); no state is kept in the secure area across sessions (any state must be explicitly stored in memory or on secondary storage).

TruztZone is implemented in ARM popular processors: Cortex-A9 and Cortex-A15 (e.g., powering platforms such as Samsung Exynos and Nvidia Tegra series) that can readily be used in laptops, smartphones, PDAs and tablets that researchers use or delpoyed in the field to extend data loggers. The description of the platform (hardware, software combination) that we are using for our prototype can be found in González and Bonnet [2013].

In our prototype, running on Trustzone, the rich area provides the data platform that rely on trusted storage primitives. Authorized storage components running in the secure area

can encrypt data and store keys in a tamper-resistant chip thus guaranteeing data confidentiality; They can verify that the encrypted data that has been stored corresponds to the data that was written, thus guaranteeing data integrity; They can replicate data stored locally on several remote instances (e.g., on the cloud), thus providing availability and durability. Note that availability and durability come at the cost of performance. Exploring this trade-off is a topic for future work. Our prototype achieves data integrity and confidentiality on top of TrustZone.

5 CONCLUSION

In this paper, we reported on our vision of a decentralized data platform for sharing ecological data sets. Much work remains to be done in order to implement this vision and to actually experiment with sharing ecological data sets. Such experiments will be crucial to validate our approach and adapt the technology to the sharing needs of various communities. The motivation and the examples in this paper are grounded in the terrestrial arctic research and monitoring community, because this work took place in the context of the INTERACT project, but we believe that other communities could benefit from our approach.

REFERENCES

Anciaux, N., Bonnet, P., Bouganim, L., Nguyen, B., Sandu Popa, I., and Pucheral, P. (2013). Trusted cells: A sea change for personal data services. *CIDR*.

González, J. and Bonnet, P. (2013). Towards an open framework leveraging a trusted execution environment. In *Cyberspace Safety and Security*. Springer.

Hey, T., Tansley, S., and Tolle, K., editors (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, Redmond, Washington.

Holder, J. and Lander, E. (2011). Sustaining environmental capital: Protecting society and the economy. PCAST Report to the President.

Park, J. and Sandhu, R. (2004). The uconabc usage control model. *ACM Trans. Inf. Syst. Secur.*, 7(1):128–174.

Parsons, M., de Bruin, T., Tomlinson, S., Campbell, H., Godoy, O., and Leclert, J. (2009). The state of polar data the ipy experience. `http://ipydis.org/documents/State_of_Polar_Data20100514_distribute.pdf`.

Porter, J. (2009). A brief history of data sharing in the u.s. long term ecological research network. *Bulletin of the Ecological Society of America*.

Wood, J. (2010). Riding the wave: How europe can gain from the rising tide of scientific data. Final report of the High Level Expert Group on Scientific Data A submission to the European Commission.