Chapter 9

Case study 4: A national solution – the UK Data Service

Matthew Woollard and Louise Corti

1. Introduction

Data Management has been succinctly defined by the editor of this volume in an earlier work as ‘an active process by which digital resources remain discoverable, accessible and intelligible over the longer term’ (Pryor, 2012, vii). Research Data Management (RDM) to the present authors is a subtly different process, which can be seen as the activities that are undertaken by a researcher or team of researchers as part of the research lifecycle, which precede but complement the activities of a data service or data service infrastructure. Theoretically, the breakpoint between the responsibilities of the two streams of activities is the transfer of data from the data producer to the host repository, but in practice this is not so black and white. In this case study we outline the key activities and services carried out by the UK Data Service that support Research Data Management as they relate to the production of research data in the social sciences. We explain how our services together with those delivered by other providers can contribute to the much-needed wider portfolio of training and professionalization in RDM in the higher education sector.

The UK Data Service began its new life in 2012 as an integrated service funded by the Economic and Social Research Council (ESRC). Aiming to create a more unified service and identity for the ESRC’s data services in the UK, it consolidates the former Economic and Social Data Service (ESDS), the Secure Data Service (SDS) and much of the data service component of the ESRC’s Census Programme. The primary aim of the Service is to provide users with access to easily discoverable and relevant data to enable and expand social and economic research. However, there are subsidiary aims which stem from this primary aim, and one of those is to inculcate better research data management practices amongst researchers and data creators. The Service is run from the UK Data Archive at the University of Essex in strong collaboration with a number of significant partners, including Mimas and the Cathie Marsh Centre for Census and Social Research at the University of Manchester along with the Geography Department at University of Southampton.

2. The structure and functions of the UK Data Service

The overall structure of the Service is based to a large extent on the functional model provided by the Open Archival Information System (OAIS) reference model (ISO 14721: 2012). The model has been adapted to provide a robust foundation for the existing and future needs of Service. Figure 9.1 shows the major areas of activity across the whole Service: it simplifies an even more complex visualisation showing the major activities in scope, and demonstrates the considerable number of interactions and dependencies within the Service. The Figure also highlights a semantic issue between the worlds of digital preservation and research management, where both communities use the term ‘data management’ to mean different things.

Figure 9.1 – Generalised scheme of the main activities of the UK Data Service

(Reproduced courtesy of the University of Essex)

The OAIS reference model is a helpful model for the delivery of digital preservation services, but for services which are not expected to provide permanent access to digital files it may be rather too complex. The key consideration in whether or not a service should follow this model is whether or not access is to be provided to digital assets on a ‘permanent’ basis, or only for the anticipated lifespan of software in which the digital files were passed into the custody of the repository. Regardless of the complexity or holistic fit of the OAIS model, it provides a very useful approach to managing the workflows of digital assets in repositories, from ingesting a Submission Information Package (SIP) of data through to the ultimate creation of a Dissemination Information Package (DIP).

As a distributed organisation, the UK Data Service is structured around key functional areas which have interlocking activities. These are explored in more detail in the subsections that follow. The Service provides outreach and capacity building across all of these functions, in particular at pre-ingest and ingest stages, which aim to instil best practices in RDM so that research data is offered for deposit in the best possible shape.

2.1 Pre-ingest

Pre-ingest activities consist of all those activities which take place before data is formally introduced into the Service. For the Service there are two key pre-ingest functions: Producer Support and Training; and Collections Development. These functions are likely to be the most critical for other organisations which offer or plan to offer research data management services.

Producer Support and Training: The Service works with a wide range of data producers, mostly government departments, international agencies, research centres and groups, and individual researchers to ensure that the benefits of best data management practices are accrued to the producers and the Service itself. The Producer Support function offers data producers guidance and advice on those elements of the research data lifecycle which precede the formal ‘deposit’ or transfer of data to the Service or another repository. This includes consideration of all data-related aspects of planning research, from costing data activities and establishing roles and responsibilities of key players through to the use of shared protocols in data collection; formatting, organising and storing the data; quality control; validation; documentation and contextualisation. Particular attention is paid to the areas that are most likely to result in precluding data sharing, namely, ethical and legal considerations, including consent and rights management. The overall approach is to explain the responsibilities of researchers to create high-quality shareable data resources and the benefits that good data management and sharing can accrue to the researcher.

The Service has worked very closely with the ESRC since 1995 to develop and implement its *Research Data Policy* (ESRC, 2013), providing guidance for grant applicants, award holders and grant peer reviewers. We have also guided the Medical Research Council in creating its data policy. Finally, the Service has been a key player in the Jisc Managing Research Data Programme (1) in the form of advice, peer review and running dedicated projects. This programme has provided a timely coordination of effort, injected funding and engendered great collaborative spirit across UK Higher Education Institutions (HEIs) to manage local research data assets. In turn this has demonstrated the pivotal role of data centres, such as the UK Data Service, in utilising their domain expertise to help develop a more unified landscape for RDM services.

Collections Development: This function ensures that the most relevant, highest impact data are selected for ingest, and actively identifies and negotiates for new data. The Service's track record in negotiating with data owners and producers to secure access to critical data sources and agree rights and licensing issues—which can be complex and prolonged—means we are able to maximise the alignment between the needs of users and data owners. A Collections Development Policy (UK Data Service: 2013) for the Service has been designed to maximise the value of the data brought into the Service's collection, since it is not possible to acquire or ingest all the data which may have value to the ESRC's target research communities or which need to be kept for some period of time. Any policy must be robust and implementable, yet flexible enough to accommodate changes in scope or direction resulting from external drivers, high-level policy changes, or indeed user demand. Our policy is supplemented by an internal cross-functional Collections Appraisal Group which ensures that the policy is implemented appropriately.

As outlined in Chapter 5, any organisation with an obligation to hold research data should design and implement a selection policy that sets out the key parameters of what will be kept, under what mandate, for how long and with what level of descriptive metadata. Such a statement provides clear boundaries for these activities, and should be backed with higher-level buy-in, support and relevant resources. At the UK Data Service, once a data collection has been though its appraisal process, a 'processing plan' is created that sets out licensing matters, how much value to add, and appropriate access (i.e. what delivery mechanisms and under what terms and conditions). The Service has a valuable role to play in training new data archives and Institutional Repositories in this bespoke appraisal process as it relates to the handling of social science-type data.

2.2 Ingest

Ingest is the activity by which data are prepared for long-term access. The activities carried out in this process range from error and integrity checking, ensuring appropriate levels of anonymisation and checking confidentiality, compiling user documentation, cataloguing and indexing, and preparing data for preservation and access. Much of the metadata created during the ingest process are used as preservation-level metadata. The skills involved in the ingest process are probably the most diverse in a single function, which is why we place a particular emphasis on their importance, especially surrounding the ingest processes for large-scale complex surveys. The key activity within this function is to ensure that there are versions of deposited data which are not software-dependent, and will be able to be migrated by the Service to more up-to-date formats when required. Institutional Repositories will need to design or adapt processes to ensure that the data ingested into their systems are usable in the longer-term, while maintaining their authenticity. This includes handling versioning of data and metadata in a robust way.

For many Institutional Repositories the ingest process is likely to be less discriminating, but the lengthy experience of the host organisations of the UK Data Service has shown that maximising ingest processing efforts as early as possible in the process of digital curation ensures that long-term access is available at a lower total cost.

2.3 Data Management

The OAIS reference model uses the concept of Data Management to cover a number of activities relating to the internal management of data. For the functional model of the UK Data Service we have elected to use the terms ‘Archival Storage’ and ‘Technical Services’ to cover most of these activities. Technical Services, however, reach beyond the Data Management activities defined by the OAIS.

Archival Storage: The Archival Storage function of a data service ensures that data are *available* in the long term. (In contrast, the *usability* of the data depends to a large extent on activities undertaken by the ingest and preservation functions.) The Service provides persistent data access, maintaining and allowing for the reuse and citation of every major version of data. Changes in data files in our custody may be necessary, either for the purposes of augmenting documentation, adding new waves to longitudinal surveys or simply correcting previously undetected errors in the data. Ensuring the logical integrity of files over time is also important, ensuring that users of data are able to validate previous research with the same data files as used in that research. The Archival Storage function also ensures that the various versions of files which we hold in our preservation system are the same as each other, providing for data integrity. The UK Data Archive (as one of the host organisations of the UK Data Service) has assessed itself against the Data Seal of Approval (2) and is continuing to play an instrumental role in helping other national data archives in Europe to attain this benchmark. It is vital that Institutional Repositories consider what level of archival status they are seeking to attain; the term ‘Trusted Digital Repository’ should be used with care and caution, as it denotes the implementation of standards at the higher end of archival practice.

Technical Services: These are closely linked with internal infrastructural activities and the overall web services. Technical Services have a profound effect on all aspects of the data service infrastructure. Almost all activities within the Service have some technical aspect, though many need human operation from the specific function. Technical Services provide the protocols for a single technical infrastructure for the Service, including internal access control mechanisms (essential to ensure integrity and provenance of data and retaining compliance with ISO 27001), preservation metadata, user identity and user access management. The last two points are particularly valid for Institutional Repositories which may have an obligation to make certain research council-funded data available for reuse, but also prevent access to other materials to which the public may not have the right to access (most likely for rights or commercially sensitive reasons).

Technical activities are increasingly managed in a single integrated workflow, for example covering a collection negotiations database, data transfer mechanisms and tools for data producers. Such seamless workflows can help to reduce both the information required from data producers and the manual activities of the data ingest team. Changes to data collections need to be fully documented in a generic and automated way, and the creation of explicit preservation-level metadata is essential. These activities can help to streamline the underlying work of any data service, so that interlocking functions are able to work closely with each other and prevent redundancy of effort.

The UK Data Service and its predecessors have been around for some 45 years, making it challenging to rebuild internal 'legacy' systems without affecting the services provided themselves. For any repository dealing with data it is important to plan as far into the future as is possible, since changes in the culture of data sharing, in access conditions and delivery mechanisms, and in the requirements of others are likely to change over time. Anticipating these changes is a *sine qua non* of any service or institutional repository wishing to remain efficient, up-to-date, and meaningful in a research world which is being increasingly driven by metrics and proof of impact.

To this end the Service has been proactive in developing access control mechanisms that better meet the requirements of the Open Data agenda, while protecting data which has the potential to be disclosive, especially when linked with other data. Our development program has included a unified user interface for internal management information, part of which can provide valuable externally-facing self-service reporting.

2.4 Access

Increasing access to, and usage of, data are high priorities for a funded data service or repository. While making research data and related research outputs available is important, ensuring they are easily findable and accessible regardless of where they may be physically located is perhaps even more critical. Seamless resource discovery is the ultimate goal for any web-based resource provider. In addition, it is vital that the existence of materials is known and documented, even if they are not accessible for certain reasons such as confidentiality, rights or other restrictions. In these cases the provision of relevant descriptive metadata is essential. Part of the rationale for the UK Data Service is to integrate the data delivery systems for data which are generally accessible and ‘secure data’, provided by the (former) Secure Data Service. Owing to the sensitive nature of the data supplied via secure access, and the formidable restrictions to access, it is not possible to harmonise these data delivery systems. However, it has been possible to ensure that the metadata catalogues for both access systems are cross-searchable. Metadata have been harvested from the ESRC Data Store—our self-deposit system for ESRC funded-researchers—to populate Discover (3), the Service’s unified catalogue. Fields have been mapped where metadata differed slightly. This allows effective cross-searching across resources previously held in different catalogues and provides a step towards enabling unified resource discovery for social science data in the UK. The Service has been contributing to national efforts to specify and agree core metadata for data collections. Our own desire is to see a scenario whereby users can easily locate ESRC-supported research data regardless of where it is held.

Much of the technical infrastructure of the UK Data Service is based on a Service Oriented Architecture, including the website. All of our data are delivered via the web in one way or another. It is important to keep in mind the importance of web mechanisms that will be used by repositories to provide access to data since users demand functional, intuitive, reliable and effective web services. Therefore, data must be easy to find, and the mechanisms for finding it easy to use. Institutional Repositories have exactly the same challenge. It is most likely that external users will be directed to an Institutional Repository from a reference to a research output which is embedded through an electronic citation. Users must feel confident of the status (or version) of the digital materials they find. The situation is no different for digital data; services must be able to provide proper persistent identifiers for the citation of the data and related objects they make available.

Efficient access to data resources not only depends on effective and user-friendly systems for resource discovery; for the UK Data Service, systems for user authorisation and user access management are also essential. Authorisation deals with a number of variant licence regimes; for example at the time of writing, census microdata available from the Service is limited to researchers in the UK, and some of these microdata are further limited to those with Approved Research status (4). Access management is important because it not only helps us understand how data collections are being used, but because it also allows us to provide this information back to the data producers and data owners. This helps to demonstrate the value of secondary analysis to both these groups as well as the funders of the UK Data Service. While most HEIs do not require complex authentication systems for their research output repositories, finding an appropriate method to collect management information is essential. For those data repositories that need to justify their user-oriented activities or secure more funding, such management information can become a vital lifeline.

For the UK Data Service, secure access, as opposed to managed distribution, will almost certainly remain the exception for data access, but the Service’s portfolio of ‘sensitive’ data will grow, as will the demand for access. The Service increasingly requires secure access methods to allow researchers to use personal sensitive data. While data owners recognise that these data can provide considerably more research value, they cannot legally allow researchers to use them without stringent usage conditions being in place. These secure forms of access are unlikely to be requirements for HEIs, but those running data management services in HEIs should be aware of the provisions of the Data Protection Act if they are required to provide access to data which might be considered to be personal. Many social surveys are indeed considered to include personal data. The 2012 guide on anonymisation published by the Information Commissioner’s Office provides further valuable advice (ICO, 2012). HEIs need to be aware of the variety of specialist services which are available to provide advice and other services.

At the UK Data Service, access to disclosive data is provided through a secure ‘portal’. However, it should be borne in mind that the costs of running such a secure access service are much more strongly related to the volume of demand than those of the ‘non-secure’ services and it is therefore essential to monitor usage and impact carefully. A vital part of delivering secure access services is in providing mandatory user training, appropriate analysis software, auditing and disclosure control functions for users. This stream of activity is highly labour- and cost-intensive and also relies on the Service conforming to the information security standard, ISO 27001, as well as further accreditation processes.

2.5 User Support and Training

The UK Data Service provides both generic and expert user support; the former is carried out across the entire Service and the latter is provided for data collections that fall into the core service at its outset. The vision for user training is to raise wider awareness of key data and to build capacity among four key groups of users: academic undergraduate and postgraduate students; academic research and teaching staff; non-academic users; and the non-traditional user.

Any effective service needs to provide some kind of assistance to users in finding what they require. In turn, this requires staff to have an understanding of the range of complexities of the data resources they provide access to. The Service has a team of expert staff, with strong research skills, focused on providing dedicated user support, advice and training. It is perhaps unlikely that many HEIs will have the resources to provide this level of support, so it becomes essential to provide as much relevant information within the Dissemination Information Package to allow end-users to be able to understand and use the data with confidence.

The UK Data Service makes use of a single web-based help enquiry system providing federated support for the whole Service. We have a strong vision to have a rich and growing bank of expert online resources for academic and non-academic users to access information and advice quickly and easily. These resources can take the form of pre-prepared help guides; multimedia training on a wide range of topics; a suite of thematic web pages; and a public-facing, searchable, web Q&A forum where communities of researchers can assist each other. These resources enable users to more rapidly find support that is tailored to their needs. We focus our specialist support on those collections relating to complex, large-scale surveys including longitudinal data, international macrodata and qualitative and mixed methods data.

A data service can never seek to replace more traditional academic methods of training, but can complement and supplement these avenues of education. Where appropriate, we point users to alternative sources of relevant information and training. In specialist areas of training such as that surrounding the use of sensitive data through our secure access systems, we focus on the practicalities of assessing confidentiality and ensuring disclosure control in microdata. These analysis-oriented training sessions complement the data management training we deliver for data creators; users benefit from having an appreciation of the factors involved in collecting data to make it sharable.

2.6 Communications and Impact

Although these activities are not part of the official OAIS functional model, they are a requirement for any data service infrastructure. Institutional Repositories may have a lesser requirement for these dedicated activities but there should be a clear rationale to improve the use of services provided at an institutional level. Visibility of resources, in addition to the quality of those resources, lead to increased use; increased use is a demonstrable impact measure and quality impact secures additional resources. This money-go-round is essential so that services are sustainable in the longer-term.

In the Service's Communications Plan, *Communicating for Impact* (UKDS, 2013b), we emphasise and explain our key focus on data use, data reuse and data sharing. We proactively target non-users, by undertaking discipline-specific focus groups, and by promoting ourselves more prominently with existing and potential data owners, in non-academic user networks and international audiences. We also seek to exploit the benefits of closer working with the key owner and producer stakeholders like the Government Statistical Services (GSS), national survey organisations and International Statistical Organisations like the World Bank. An internal communications team focuses on promoting our funder's impact agenda through the exploitation of the management information we collect. This function also coordinates, monitors and promotes educational and research impacts, working closely with ESRC and other data suppliers.

2.7 Preservation Planning

The Service ensures that its preservation-based activities are fully grounded in international best practice, and liaises with key organisations and networks over these activities. We maintain our own standards and technology watch to ensure that the best value for money is maintained for our funders. In addition, we provide informal training and mentoring for other international social science data archives that are setting up data preservation services.

We do not touch on management and administration of the Service in this chapter, but suffice it to say that management commitment is vital for the smooth running of the UK Data Service; likewise HEIs will need to provide dedicated resources for the management of their repositories if they are to be successful.

3. Conclusion

This overview of the UK Data Service shows how our activities, many of which are standards-based, can influence and complement research data management infrastructure in HEIs. Five interlocking factors are causing HEIs to rethink their research data management services:

- researchers' responsibilities towards their research data are starting to change;
- research funders are increasingly mandating open access to research data which they have funded, requiring relevant data management planning and practices in order to maximise transparency and accountability of all research areas;
- governments are demanding transparency in research;
- journal publishers increasingly require submission of the data upon which publications are based for peer-review;
- and the economic climate is requiring much greater reuse of data.

Together these drivers mean that researchers will need to improve, enhance and professionalise their research data management skills to meet the challenge of producing the highest quality research outputs in a responsible and efficient way, with the ability to share and reuse such outputs. These initiatives also mean that HEIs will have to step up to the mark in their activities to support long-term access to these data and to manage the ethical and security risks of their data assets. Their responsibilities will change, and investment to develop capacity will be required.

Data centres can play an invaluable contribution to building capacity in a number of areas, particularly where there are discipline-specific issues surrounding the data resources and known user-needs.. One of the UK Data Service's host institutions, the UK Data Archive, has worked on a number of projects which provide research data management advice to universities, institutional repositories and researchers. Both the UK Data Service and the UK Data Archive expect to carry out these activities into the future, and we can provide bespoke advice and capacity building to the emerging institutional data repository landscape.

4. References

4.1 Websites and notes

(1) Jisc Managing Research Data programme 2011-2013: http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/managingresearchdata.aspx
(2) Data Seal of Approval: http://datasealofapproval.org
(3) UK Data Service Discovery interface: http://discover.ukdataservice.ac.uk/
(4) Access to ONS Data: http://www.ons.gov.uk/ons/about-ons/who-we-are/services/unpublished-data/access-to-ons-data-service/index.html

4.2 Bibliography

ESRC: 2013 ESRC *Research Data Policy*. September 2010. Revised March 2013. Available at: http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf

Information Commissioner's Office: 2012 *Anonymisation: Managing data protection risk. Code of practice*. Available at: http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Practical_application/anonymisation_code.ashx

ISO 14721:2012. Space data and information transfer systems — Open archival information system (OAIS) — Reference model

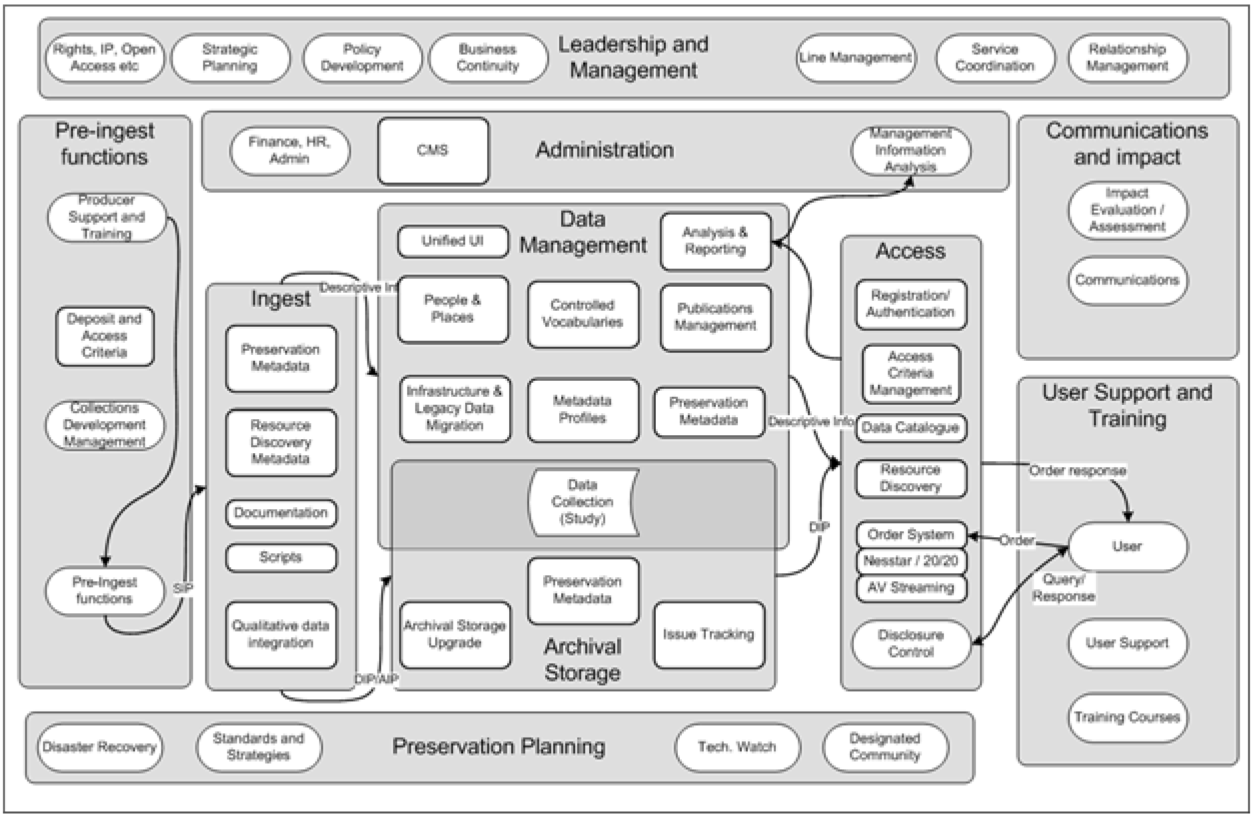
Pryor, G. (2012) Preface. In Pryor, G. (ed), *Managing Research Data*, Facet Publishing.

UK Data Service: 2013a *UK Data Service Collections Development Policy*
Available from: http://ukdataservice.ac.uk/deposit-data.aspx

UK Data Service: 2013b *Communicating for Impact*
Available from: http://ukdataservice.ac.uk/about-us/impact.aspx

FIGURES

Figure 1: Generalised scheme of the main activities of the UK Data Service


Rights, IP, Open Access etc
Strategic Planning
Policy Development
Business Continuity
Leadership and Management
Line Management
Service Coordination
Relationship Management
Pre-ingest functions
Producer Support and Training
Deposit and Access Criteria
Collections Development Management
Pre-Ingest functions
SIP
Finance, HR, Admin
CMS
Administration
Management Information Analysis
Communications and impact
Impact Evaluation / Assessment
Communications
Ingest
Descriptive Inf
Preservation Metadata
Resource Discovery Metadata
Documentation
Scripts
Qualitative data integration
DIP/AIP
Data Management
Unified UI
Analysis & Reporting
People & Places
Controlled Vocabularies
Publications Management
Infrastructure & Legacy Data Migration
Metadata Profiles
Preservation Metadata
Data Collection (Study)
Preservation Metadata
Archival Storage Upgrade
Issue Tracking
Archival Storage
Descriptive Info
DIP
Access
Registration/ Authentication
Access Criteria Management
Data Catalogue
Resource Discovery
Order System
Nesstar / 20/20
AV Streaming
Disclosure Control
User Support and Training
Order response
Order
User
Query/ Response
User Support
Training Courses
Disaster Recovery
Standards and Strategies
Preservation Planning
Tech. Watch
Designated Community