{logo?} { Community logos?}

NHS- Data Access Service (DISCOVERY)

Version 2.0
Draft

Amendment History

Version	Date	Amendment History
0.1	1/10/215	First draft for comment
2.0	20/12/2015	Generic version

Authors and Reviewers

This document must be reviewed by the following:

Name	Title / Responsibility	Date	Version

Approval

This document must be approved by the following:

Name	Title / Responsibility	Date	Version

Contents

1 Document Purpose

The purpose of this document is to describe a software service designed to enable local NHS organisations to have greater control of access to the data held within their systems.

The document has two parts, which describe:

1. Why this is necessary, what the software services are, and the anticipated benefits that are enabled by this type of service
2. A description of the technologies that make up the service

2 Introduction

2.1 Population health

Predictive analytics is an essential enabler to drive population health improvement and optimise the quality of care delivered informing both direct clinical care and third party use for planning, development and research purposes.

Achieving this personalised predictive analytical capability will depend upon access to real time clinical data, collected at the point of care.

There are a number of obstacles to data access currently. Some of these are:

1. The current architecture of healthcare IT systems represents an obstacle to delivering the required level of access to the data needed by front line clinicians, managers and patients. The architectural problems are:
    a) The systems host the patient data on behalf of a provider organisation (e.g. a practice or a Trust).Consequently they each only hold a proportion of a particular patient's data. Modern healthcare decision making often requires a view of, and analysis of, the patient's data as a whole i.e. access to data entered via a number of different systems and a number of organisations.
    b) Most of the system suppliers consistently limit access to the data in their systems, either for commercial reasons or because of limited capacity to enable access.
    c) The data held in systems are held in a variety of different formats and structures and different coding schemes, so it is very difficult (and expensive) to analyse across different systems as each system format has to be tacked for each analysis project.
2. The current data sharing models for direct care vary between systems, and the sharing models are incompatible with each other. The likely clarification of sharing issues by Caldicott early next year for both direct and indirect care will enable a common model to emerge. However, it is unlikely that the suppliers will be able to make the necessary changes at sufficient pace
3. The business case for moving to common standards such as the adoption of a single clinical terminology (SNOMED-CT) is well established. However many

suppliers will be unable to undertake this migration without sacrificing opportunities for other developments.

4. Incumbent suppliers have established development road maps and they are generally falling behind schedule with developments, these developments being increasingly complex. It is unlikely that they will have the capability to take on multi-organisational and multi-domain population health analytical requirements.
5. Current proposals to establish "fixed data sets" for commissioning purposes are unlikely to deliver what is needed. They address neither the need to develop informatics for direct clinical care of patients, nor the information needs of patients and their carers'. Put another way, there is an urgent need to be able to use a constantly changing and developing data sets without having to reformat the data stores and renegotiate with separate suppliers. Modern data analytics use data streaming to cloud based stores that can address all three requirements.
6. The current costs of data hosting are high due to the legacy architectures and legacy hardware.

2.2 Data access principles

When considering the establishment of more efficient ways of access data, the following set of fundamental information principles apply:

- Healthcare and social care providers have a duty to share information, derived from data held within their systems, with other healthcare providers and their patients that use other systems.
- Healthcare providers have the need and the right to be able to access and use their own data for their health and care business purposes.
- Patients have the right to expect that the data recorded by their healthcare professionals will be used for their care and shared with others when appropriate
- Suppliers have a duty to provide access by the data controllers, to their own data, to the extent required by the data controllers.
- The NHS has a duty to ensure that the systems in use within the NHS and social care sector are able to fulfil these needs.

The Newcastle CCIO Declaration states:

"All clinically relevant data held within supplier systems must be made available for use in any care setting, wherever and whenever required, subject to relevant Security, information governance, and consent requirements."

2.3 What DISCOVERY is

DISCOVERY is a supplier neutral data access service that is designed to enable access to data by other systems and suppliers, access being under the direct control of local NHS organisations that own and share the intellectual property of the technologies within the service.

The data access service is a set of hosted software components, operating as a set of services, which provide access (by the data controllers) to an up to date logical equivalent of patient and related data held in their care management systems.

The service is owned by the NHS organisations themselves with the running of the service either being operated in-house or contracted out.

The following things make up the software services:

1. An open source, common standard, resilient, hosted, secondary data store, holding patient clinical and organisation workflow data. The patient data is keyed by patient NHS number and by the data controller.
2. An open source, hosted software service that populates and maintains the above secondary data store in near real time, with clinical and workflow data posted from operational systems.
3. An open source, hosted software service that provides secure access to the data via a set of NHS standards based APIs, together with open standard query, controlled by an NHS level security model including role based access rights and data sharing rules.
4. An open source, hosted software service, that generates additional special purpose databases from this data store, for use by proprietary software vendors, including the main suppliers themselves.

The following types of owner NHS organisations are proposed:

a) A community member national service provider organisation either with members made up of local organisations, or trusted by them to provide the service. The national service is responsible for the hosted environment and core access support and holding of contracts for external service supply.

b) Local or regional NHS organisations, as members of the above or supporters of the above, with responsibility for local project management, local configuration for access, and local extensions to access service software and service.

In other words, a consensus based organisation managing the core on behalf of local organisations and the local organisations controlling access and extending the service in their local area. This is thus NOT a single national organisation but a series of organisations that share technical resources to the maximum extent.

The sources of the technologies for the data service are:

a) Code-4-health communities, designing and building the software, brought together into a single managed build and testing environment.

b) HSCIC and NHS England.

c) Other parties, including commercial parties, prepared to contribute open source components consistent with the architecture.

d) New cloud style hosting suppliers willing to match commercial cloud level hardware prices.

The sources for setting requirements for accessing the data are:

a) Lead clinicians and clinical informaticians leading both the local and national organisations taking an active role in the design of the services.

b) NHS England or HSCIC programs involved in clinical data access projects (such as commissioning data service, GPES, SCR)

The governance and security rules around the service include:

a) The primary data controller retains control over their data within the service at all times. In other words, the level of control remains identical to that covering their main system's data store.

b) There is no requirement for additional patient consent as this is operating on behalf of the data controllers in the same way that their current systems do.

c) Access to the data store is secured and managed to the same Information Governance levels that apply to current nationally hosted clinical solutions such as PAS systems and GP systems.

2.4 What DISCOVERY is not

- It is not an Integrated Digital Care Record (IDCR). It is an alternative source of data that can be used within one.
- It is not a Health Information Exchange (HIE). It is an alternative source of data that can be used by one.
- It is not a replacement for health records. It relies on the operational health records to provide the data to populate it.
- It is not a portal. Portals may be used to access it.
- It is not a common record. Whilst it is patient centred (keyed by NHS number), it does not, for example, remove duplications or inconsistencies. It is however, an alternative source for data to populate a common record if there was a need to create one.
- It is not a Personal Health Record. It can be used to represent an alternative source for a significant proportion of the NHS data needed by personal health records.

2.5 Differences from other similar data initiatives

- Care.Data is a pseudonymised limited data set used for a variety of research and commissioning activities. It requires changes to data controller status of the data once extracted and has a different set of customer use cases. The DISCOVERY service could be used as an alternative technical source for it.
- GPES is the GP extraction service that only extracts limited amounts of data. It processes only a few queries per year. As a horizontally scalable source of data, the DISCOVERY service would process thousands of queries.
- Commercial health record repositories could be used for this purpose. However, as currently configured, as they are used for common shared records, the data is pre-filtered by data sharing agreements before extract. They are commercially owned. The DISCOVERY service has no license fees and the data controllers own the IP.
- Local extract data stores for analysis. These could be extended to meet the requirements. As currently configured they are proprietary, one way only and used only in a pseudonymised manner for indirect care. DISCOVERY could be used as a source for pseudonymised data for this purpose also.
- Specialist data extraction companies represent commercial equivalents to the extract elements of this service. They are commercially owned. DISCOVERY is owned by provider organisations.

3 Enablement and Benefits

There are a set of enablers derived from this service and a set of healthcare related benefits.

3.1 Enablers

By creating an open standards and open source based architecture for both storage and access via standard interfaces the following are enabled:

- Reduction in dependency on current system supplier's capabilities and willingness to interoperate
- Allows a more rapid move to open standards without disruption to live systems by demonstrating and resolving all of the data upgrade issues in a transparent manner
- Getting new data access APIs quickly to market, allowing new and existing software systems to provide earlier access to new functionality
- Easier upgrade path for the next generation of systems as the open standard model has been proven and tested
- Remove the need for locked-in partner programmes or GP-SOC pairing and encourages new entrants to the market, led by demand
- The benefits of open source development and sharing leads to a wider pool of enhancement capability via the open source community, including local extensions, well beyond that achieved by proprietary software
- The user, as data controller and service owner can direct the developments they need
- As it is owned and managed by new types of NHS organisations this means that the service is an NHS asset and not a commercial supplier asset

The creation of patient centred records (independent of the main clinical applications) using data sourced from different systems, together with appropriate data sharing agreements and common open standard records will:-

- Enable the creation of a common query to apply across a population. As the data store uses an open standard format, a standardised query approach can be used
- Enables applications to be designed, including patient facing applications that allow multiple record sources to be used within the application with a single API call. As access to the data uses standard APIs, applications can be built in the knowledge that they will also operate on other data sources as the API standards become used
- Allow the creation of additional storage resource to reduce the load on the primary system's databases and enable significantly greater volumes of query and subsequent access to the data

This then enables applications to perform the following types of activity

- Advanced analytic applications can be used to identify or stratify risk for patients who need interventions; providing the sort of analytics that is well beyond the simple QOF style rules engines currently provided.

- The ability to analyse patient's variances from, or adherence to, relevant care pathways that are implemented as part of the new models of care services, in order to improve quality of care provision and better outcomes
- Advanced dash-boarding applications can be used to determine current and past status of the business, e.g. appointments take up, workload distribution as well as the ability to project trends in activity
- The ability to share clinical record subsets, with other organisations, by applying a common data sharing agreement in a way that resolves the conflicting supplier system sharing models
- As an alternative source of patient data for patient applications or PHRs it provides a platform to enable many more innovative patient centred applications, to operate with the patient's data, treated as a whole rather than a set of disconnected data sets
- The ability to provide an alternative source of data to be extracted for the commissioning data service

3.2 Benefits

The following are a few examples of health business benefits that can accrue as a result of being able to bring in new applications to operate on the linked data:

Unscheduled care

Patients can access unscheduled care from a variety of services including network or locality hubs, GP out of hour's services, walk in centres and Accident & Emergency Departments.

An integrated data service would analyse individual patient journeys through the entire health system, segmenting patients into categories in order to match evidenced based interventions that can favourably alter patterns of behaviour or medical interventions.

Chronic kidney disease

Population interventions which delay progression to CKD stages 4 & 5 are key to reducing the progressive rise in dialysis accruals. Combining laboratory data from primary and secondary care would enable population tracking to ensure diagnostic coding and support prompts to enable the effective primary care management of early CKD.

Gestational diabetes

Gestational diabetes is recorded as a problem in the hospital EHR. It is recognised that a proportion of these women will develop Type 2 diabetes related to weight gain or subsequent pregnancies.

A recent study has shown that in one area only 50% of these data items are transferred to GP clinical systems. Consequently inadequate follow up occurs in general practice.

A data service with population health analytics would ensure that these women are tracked and provided with high quality pre-diabetes and diabetes care

These IT processes go hand in hand with transformation in the social organisation of care in both primary and secondary care settings, placing unprecedented information at the clinical coal face with potential for better patient self-monitoring and engagement.

Cardiovascular pathway- Atrial fibrillation

While the pathway can be described in detail the primary care, systems are unable to link the individual patient journey from anticoagulation to adverse outcome in terms of stroke or bleeds. (This functionality is simply not possible even in bespoke national registers such as the Sentinel Stroke National Audit Programme SSNAP).

By linking the primary and secondary care pathway, this enables individual linkage of process and outcome data. This can use routinely collected data to show how many people on anticoagulants have a stroke or bleed which will inform risk profiles and optimal drug prescription. For example, flagging those patient prescribed inappropriate medication, with poor anticoagulant control or at risk by virtue of poor renal function

4 Service description

4.1 The current architecture

Figure 1 illustrates the ownership, access control, and make up of a typical set of systems in the NHS. The components are described in more detail following this:

Figure 1

As a rule, most systems have their data held in a hosted environment in secure data centres connected by N3 to the organisational networks. For example, over 90% of practices have all of their data hosted in three hosted instances (EMIS Web Systm 1, and In Practice). These are designed to be resilient with offline read only data stores duplicated in a data centre and duplicated across two data centres.

Hospital systems, are increasingly, hosted in Trust owned or Trust leased hosting centres. Examples are Lorenzo, Millenium and openMAxims.

Applications that are hosted have a separation between the application and the data store. The application and data store communicate with each other via layers of internal software connection methods known as “interfaces” or “APIs”. There are usually between 3 and 5 layered tiers between the user interface and the data store.

In addition to the application, the systems also support a communication tier that communicates between the data store and the outside world. Most of the systems name this tier. For example, EMIS uses 'EMIS Connect' and Systm1 uses 'Systm Connect'. These applications handle communication with third party communication modules including for example the NHS spine TMS, DTS, and commercial partners.

Most systems also have a communications connection within the local organisational network but these are limited to systems installed on the same LAN or device.

In addition to this architecture each of the systems normally has the ability to "bulk extract" some of their patient data into a large file. This is normally an initial upload followed by an "incremental extract", perhaps every day but mostly monthly.

Access to the data by remote organisations is thus constrained by three tiers:

1. The application itself
2. The system's communication tier
3. The system's bulk extract service

All these are controlled and maintained by the main supplier.

Figure 1 (page 10) illustrates one system (system E) that is hosting data for multiple practices operating through a single application and another (system S) with only one practice for brevity. Other systems communicating by the APIs are included.

A market is established for commercial bulk extract specialists who provide data either to their own analysis applications or data to other analysis applications. Examples of these are Apollo and BMJ Informatica.

Extracts are generally bulk followed by monthly or weekly increments, although in some cases are daily.

In this model the user is dependent on both the extractor supplier and their main supplier for access to their data either directly or indirectly through the analysis application. Different analysis application suppliers are dependent on different extractor suppliers.

In practice, this restricts the number of extractors and also restricts the number of analysis application providers, and thus prices reflect the supply and demand mismatch.

Most current specialist extractors only extract the amount of data specifically required for analysis purposes, which is not sufficient for use as a care record view. Those who do use it for record access, mostly provide only a summary extract as, in this case, the record is only used for sharing purposes. It is important to note that control of the data is transferred during this process.

If the end user is from another organisation, other than the primary data controller, the current model requires that data sharing agreements are in place prior to any data extract. These are shown in blue in figure 1 as operating within the supplier-hosted environment.

This means that only certain data fields are extracted from the main system, those fields and codes being limited to those covered by data sharing agreement. Therefore, the data controller sometimes has less data for analysis than they have in their own host system.

When further extract projects are required (and they change regularly), and they need different data fields and thus changes to the data sharing agreements, this requires the same and different data to be extracted again from the main systems.

This is unwieldy as it creates a dual dependency on the commercial relationship with the main and extractor commercial suppliers and requires amended data sharing agreements prior to repeated bulk extract.

4.2 Proposed architecture

Figure 2 shows the proposed architecture. This proposed approach would enable data controllers to substitute commercial extractor services with a single vendor neutral extractor service. The service will extract data only once from the main system and will, subject to IG standards, host the data for a variety of uses as determined by the data controller. This will ensure that:

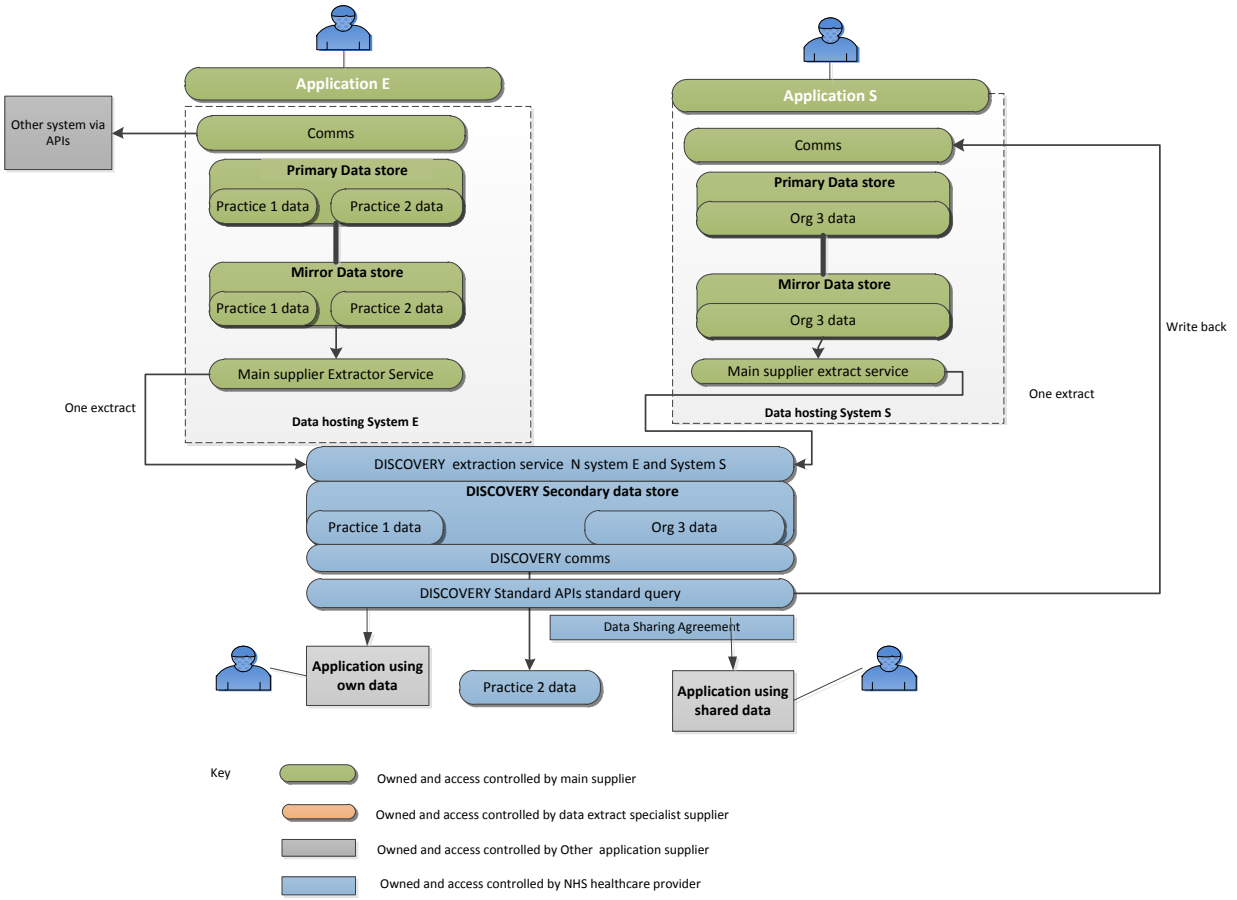
a) The data controller is able to control further access to their data without recourse to their main supplier and specialist extractors

b) It is possible to increase the number of fields in the extracted data i.e. a full set of fields to support detailed analysis and record construction

c) Data is bulk extracted initially and incrementally extracted as per current practices. However, it is envisaged that the incremental extract can occur in near real time (see technical sections) as this is done only once per main system.

d) New data generated from other applications can be written back

Fig 2



The main differences between the architectures in figure 1 and figure 2 are:

1. A secondary store is created and is both owned and controlled by the data controller's organisation.

2. The data is extracted once for many purposes rather than many times for many purposes.
3. Data sharing agreements are applied outside the suppliers systems in DISCOVERY itself.
4. Data can be "written back" to the main systems. An example may be a complex risk score that is not available in the primary system, posted back to the primary system and used for patient management purposes.

4.3 Proposed standards

The following open standards apply to the DISCOVERY structures

1. The Contributing data from source systems will contain coded data, generally in either READ, ICD, Snomed/ DM+D or other. These codes will be stored as Snomed-CD (preserving the original code and term in addition)
2. The format of the data will be an emerging record format using the PRSB headings. Likelihood will be that it is held either as openEHR or FHIR
3. The APIs for subsequent data access will be the NHS England/ HSCIC standard APIs
4. External message content format would be CDA or FHIR (depending on the API)
5. The standard query format would be a standard such as SQL
6. Authentication and security would follow the NHS England/HSCIC standards for system to system communication (e.g. as per ITK and Spine TMS) It is not expected that DISCOVERY would be responsible for end user authentication as this would be managed via locally selected, authenticated and trusted applications.

In addition to standards, the service will need to support a variety of approaches for systems that have not been upgraded to the standards. For example:

1. CSV outputs to third party applications
2. Business oriented simply data APIs mapped from the open standard APIs

4.4 Service Technical components

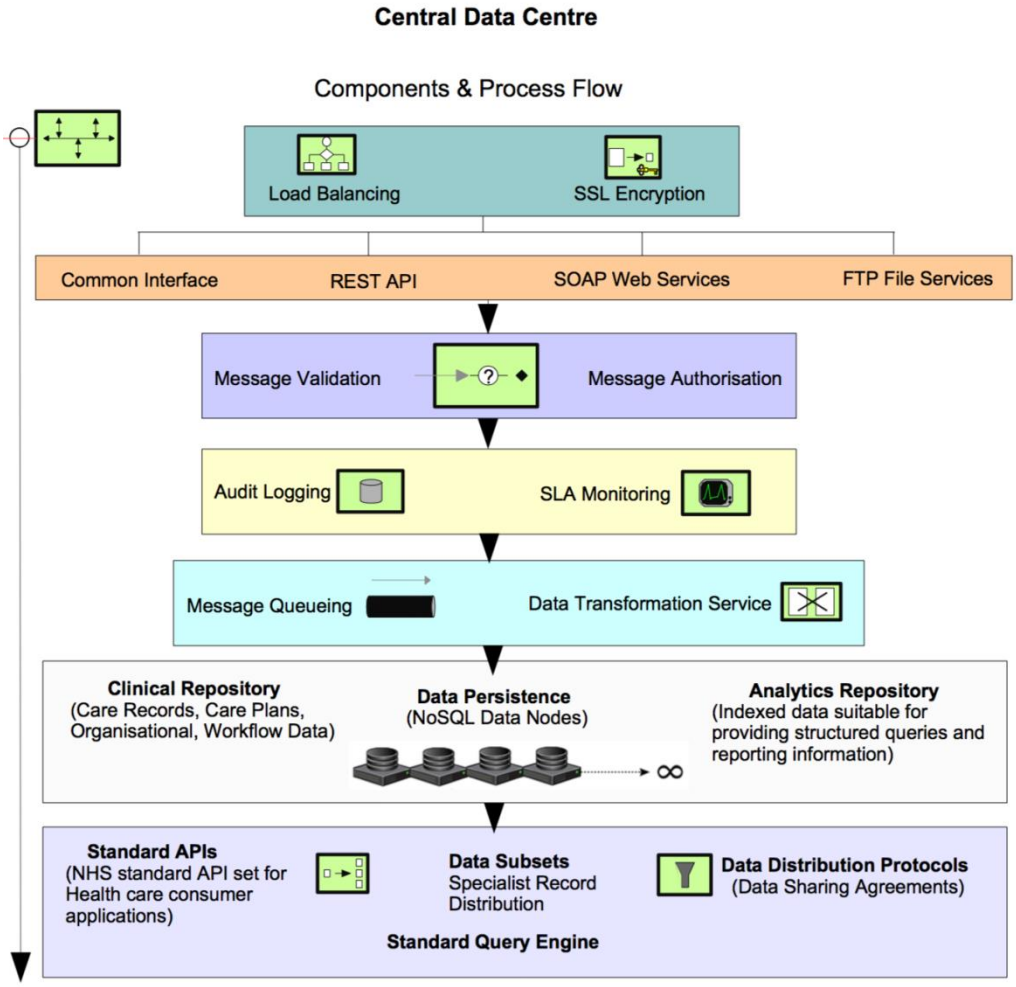
This section summarises the relevant technical components and architecture of the DISCOVERY service.

The service comprises of several categories of technical components:

- Primary systems data extraction and upload service
- Service load balancing, high availability and scaling
- Service communication interfaces
- Authentication and security
- File and message processing engine
    - Inbound data validation
    - Inbound data queues
    - Inbound data audit
    - Transformation and mapping services
- Secondary data store
- Data sharing agreements / protocols
- Service monitoring and deployment
- Standard APIs for consumer applications
- Query engine and standard query languages for analysis applications

Figure 3 illustrates some of the key components and process flows within the DISCOVERY service.

Figure 3.



4.4.1 Primary systems data extraction and upload service

This service is responsible for extracting the data from the main systems in a timely manner.

It operates within the primary system's hosted environment.

It uses the concept of a "bulk and incremental upload" which is an extract of the full coded and text record from a single patient or a batch of patients. A Bulk upload is divided into three concepts:

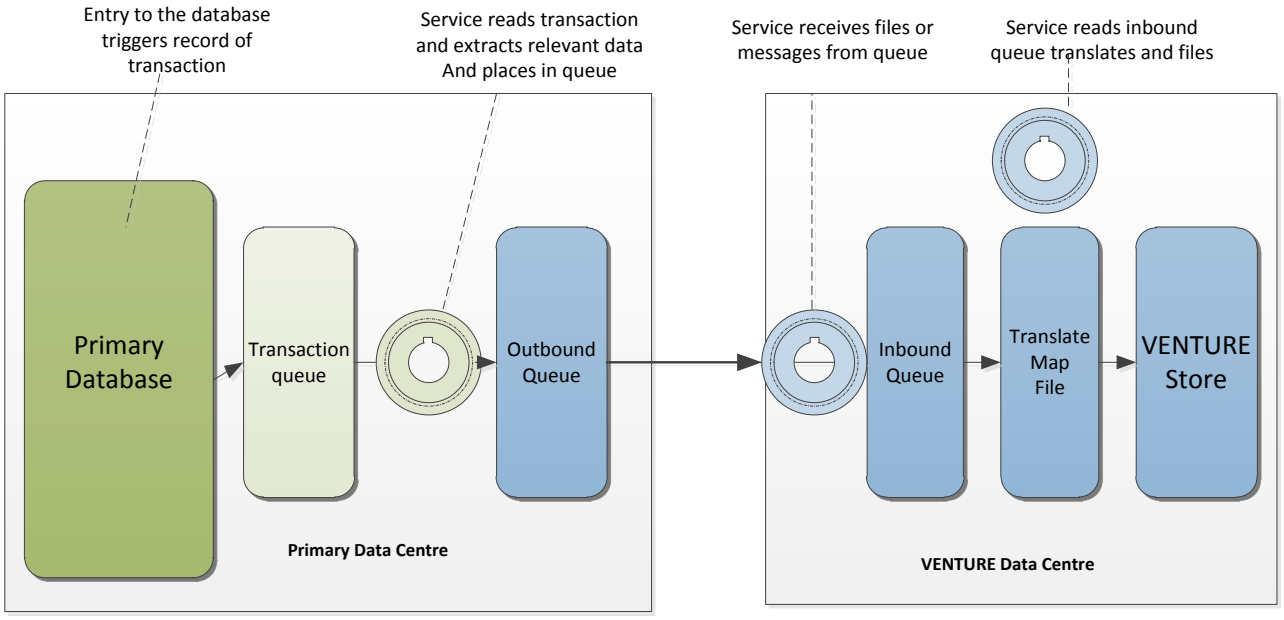
1. Initial upload, covers qualifying data for all patients that qualify for data upload and workflow items that qualify for upload.
2. New uploads, covering the records of patients that qualify for data upload but were not in the initial upload list.
3. Repeat uploads, covering repeated uploads of selected patients or re-runs of initial uploads.

Incremental uploads are triggered by changes to the data, including new data, deletes and edits.

Incremental uploads should be as close to real time as possible, determined by the end use case. (Current systems support daily but based on a transaction queue so could be brought closer to 10-30 minutes).

The architecture of the data extraction and upload service is illustrated in figure 4

Figure 4



Modern hosted systems use modern databases such as Oracle or MSSQL that have triggers associated with update and delete transactions. These can be used to create a log of transactions (transaction queue).

The upload service reads the transactions, generates the data to be posted to the service, and then ideally places it in an outbound queue. This is then transmitted to the inbound queue within the DISCOVERY hosted environment.

Data can be exported from systems in one of two forms:

1. Flat file CSV format with documented cross links
2. Structured messages, one per transaction (full patient record or transactional record).

4.4.2 Service load balancing, high availability and scaling

The DISCOVERY service and data store is ultra-resilient and massively scalable using the concept of “horizontal” scalability.

High availability is achieved by the elimination of single points of failure. This means adding redundancy to the system so that failure of a component does not mean failure of the entire system.

Failures of system components are detected as they occur, and all errors are immediately alerted to service personnel who have the monitoring and administration tools required to diagnose and rectify system errors.

The service communication endpoints are protected from processor overload by high-availability pairs of load-balancing nodes. These load balancers distribute the workload across multiple computing resources, called clusters.

This initial layer of computer servers performs a number of processing tasks before the data is committed to the data stores. This layer is often referred to as the service or application tier and the computer servers are called “nodes”. These processing tasks include authentication of messages, data validation, queuing of data, data auditing, translation and mapping of data etc.

The server clustering approach connects any number of readily available computing nodes via a fast local area network inside the data centre. The activities of the computing nodes are orchestrated by "clustering middleware", a software layer that sits on top of the nodes and allows the software to treat the cluster as one large cohesive computing unit.

The primary scaling architecture of the service is often referred to as “horizontal scaling”. Both the service layer, described above, and the data store layer, described next, must use horizontal scale-out methods (or elastic scalability).

To scale horizontally (or scale-out) means to add more nodes (computers) to a system as and when increased load demands. As computer prices have dropped and performance continues to increase, high-performance and low-cost "commodity" servers can be used for tasks that once would have required supercomputers.

DISCOVERY is designed to allow the configuration of hundreds of smaller server computers in a cluster to obtain aggregate computing power.

Each server computer is modest in specification, and highly cost efficient within a modern cloud based data centre environment.

As each service node performs the exact same set of tasks, new servers are added into the cluster to increase the shared capacity of the overall workload, and faulty or redundant servers can be removed from the cluster without affecting the system, and without any loss of service availability.

The data store must be equally (and in many ways even more) scalable and resilient than the service layer.

The integral core data store technology within DISCOVERY uses scalable database technologies called NoSQL, often referred to as ‘big data’.

Relational databases are expensive and difficult to scale-out to extreme sizes. They have particular problems processing high volumes of new data and are more suited to the analytical purposes (see below).

The DISCOVERY service supports NoSQL data store structures that are optimised for accepting new data as well as relational database structures for subsequent analysis. The service uses open-source industry leading NoSQL database technology in order to achieve high-availability and ultra-scalability within the data store layer.

The NoSQL technology has the following main characteristics:

- Elastic scalability
- Always on architecture
- Fast linear-scale performance
- Flexible data storage
- Easy data distribution
- Operational simplicity
- Transaction support

In the database design, cluster all nodes are equal. There are no ‘master’ nodes and therefore no single point of failure offering true continuous availability and uptime. All nodes communicate with each other equally. This means that if the first node that the

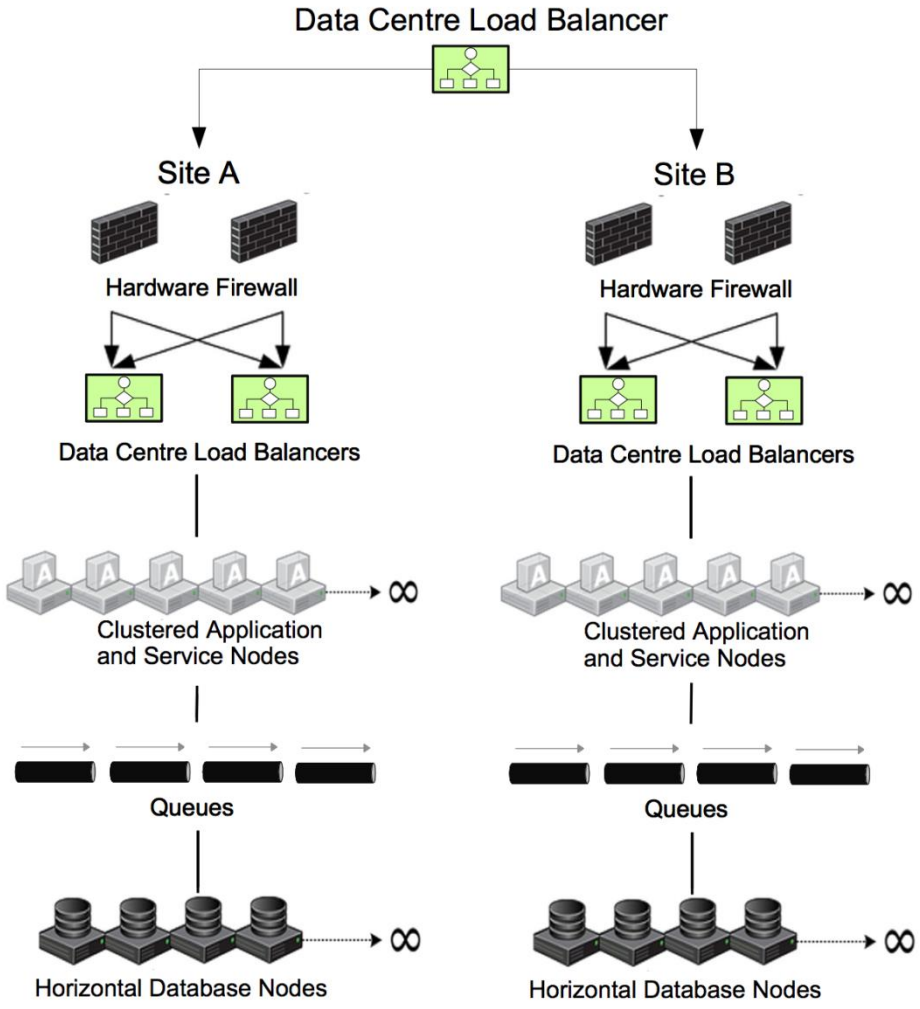
service components are connected to goes down, the database technology knows about the other nodes and can connect to one of them instead.

This built-for-scale database architecture means that it can be capable of handling huge amounts of data and thousands of concurrent users or operations per second - even across multiple data centres.

When further load and data is required, technicians simply add new nodes to an existing cluster without having to take it down. When a database node is faulty or needs taking down for maintenance it can simply be removed from the cluster with no effect on the system. This is made possible as data is automatically replicated a number of times across nodes and disks, with the same data always available on multiple nodes.

A simple diagrammatic representation of load balancing and server node scaling is shown in figure 5.

Figure 5.



4.4.3 Service communication interfaces

DISCOVERY uses a number of communication layers, each containing many interfaces or “APIs”. An interface is a software component that provides a function with inputs and outputs. Other software applications can make calls on these interface functions, passing in data or instructions (called input parameters), and receiving outputs (data or response acknowledgements).

The first type of interfaces that DISCOVERY exposes is a set of ‘data receivers’. These functions are used as part of the extraction and upload service described in section 4.4.1.

There are two main types of interfaces used to extract data from the primary systems.

One set of interfaces can handle files (such as CSV files). These interfaces are referred to a File Transfer Protocols, andare used to securely fetch or receive data files ready to be translated into structured data records.

The second set of interfaces handle structured messages (such as FHIR, openEHR, OpenHR, EmisOpen, HL7, CDA etc.). These interfaces receive structured data wrapped in authentication envelopes (headers). DISCOVERY receives data in many different structured formats via these interfaces, ready for the translation and transformation processes that get the data ready for filing into structured database records.

These message capable interfaces are orchestrated by modern communication protocols such as REST and SOAP. All messages within DISCOVERY conform to industry standard authentication and security rules and NHS compatibility toolkits such as ITK.

Once DISCOVERY has received the primary system data it is then passed on to the next layer in the message processing chain.

4.4.4 Connection Authentication and security

All communication via networks into and out of the DISCOVERY service are protected by the highest level of industry recognised security and encryption services.

DISCOVERY interfaces are protected by industry standard Transport Layer Security (TLS). This is a cryptographic protocol designed to provide communications security over a computer network. The primary goal of the TLS protocol is to provide privacy and data integrity between two communicating computer applications.

Identification digital certificates containing the server name, the trusted certificate authority (CA) and the server's public encryption key should be used in addition.

DISCOVERY should use TLS 1.2 with SHA-256 cryptographic hash functions.

4.4.5 File and message processing engine

Once data has been received by the DISCOVERY service it needs to be processed through a number of different channels before it can be filed (persisted) into the data stores.

Primary data needs to be validated, queued for processing, audited and translated into a standard format.

4.4.5.1 Inbound data validation

Each file or message received, by DISCOVERY, from a primary system must be initially validated to ensure that the data received is the data that was expected and conforms to known rules.

If inbound data is not from a known source, or in an unrecognised format, or is not consistent with a data sharing rule (see below) then it cannot be further processed by the service, and a response is sent to the sender of the data, informing of exactly why the message or file has been rejected. All response messages have codes, and these industry standard codes have accompanying messages with full descriptions of why a rejection has taken place.

If a message or data file passes this basic validation check, then the message is queued for further processing, and an acknowledgement message is returned to the sender informing of successful message receipt.

This will not be the only acknowledgement message code sent to the message sender. As messages are process asynchronously (for reliability and scale reasons), a second acknowledgement message will be sent back to the sender once a message has been successfully processed (normally on safe filing into the data store).

4.4.5.2 Inbound data queues

All data received from primary systems is processed asynchronously, and will therefore end up in a series of queues. Queues are also a reliable technology option for increasing the workload capabilities of message processing systems, as many processes can concurrently read messages from queues.

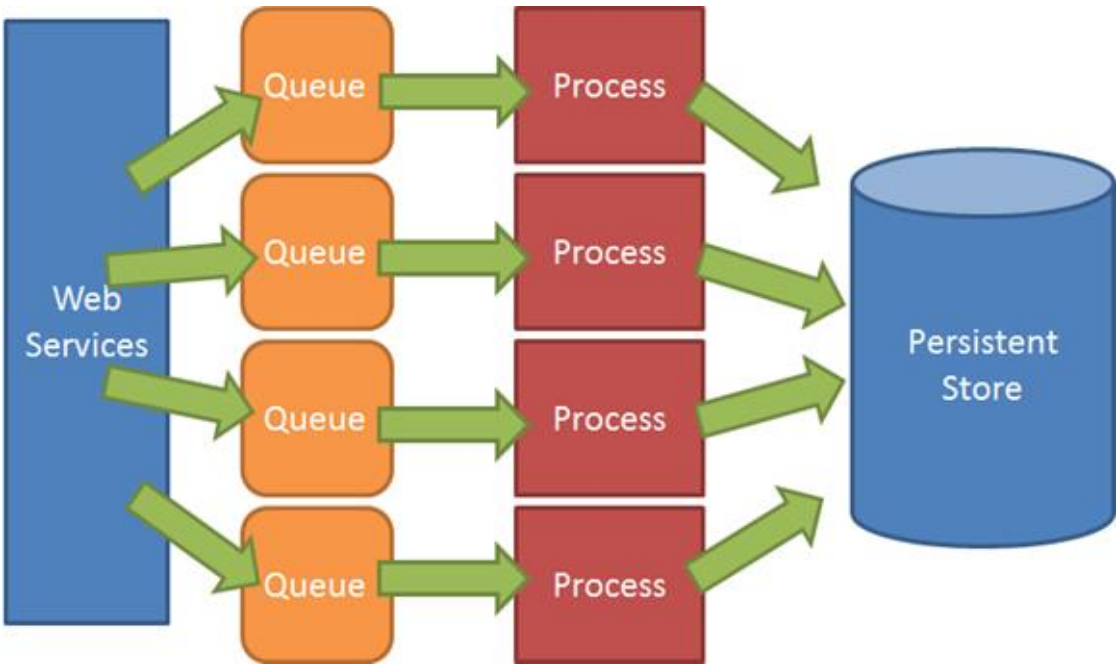
Queues are temporary in-memory / data-cached files which store messages for further processing. Messages are normally processed through queues in a first-in-first-out basis. The system will have further complexity than this with a series of different types of queues for different types of purpose; for example - priority message queues, bulk-processing queues, transactional queues, and even organisational specific queues.

The queuing technology used in DISCOVERY must be industry-leading open-source. It must have the following key characteristics:

- Reliability
- Flexible routing
- Clustering capabilities
- High-availability
- Multi-protocol capable
- Management UI
- Tracing capabilities

Figure 6 illustrates how multiple queues can be used to temporarily store messages before further processes can translate and store the data contained within messages into persistent database record formats.

Figure 6.



4.4.5.3 Inbound data audit

All inbound files and messages that are received by DISCOVERY must be audited for security and administration purposes. Raw messages are logged in an audit log that can be used to trace data discrepancies and errors down the line. It also allows a comparison of data pre and post data transformation. If there is an error or inaccuracy in the data transformation process (see next section) then the mapping can be verified by looking at the differences between the before and after versions of records.

Audit is also a safe way of tracing events in a system, and audit provides a transaction log of events in case these events need to be scrutinised later for Information Governance or other diagnostic reasons.

Audit logs will be maintained for a period no longer than is required to maintain a safe and well-governed data service. If data is permanently removed from the data stores via data controller request, then all accompanying relevant data will also be removed from the audit logs.

4.4.5.4 Transformation and mapping services

Data transformation (translation and mapping) must be a key part of the DISCOVERY service.

Different systems store data in many different formats. All formats tend to be well structured and nationally recognised. However if the DISCOVERY service was to store all data from different primary system in the original formats, then it would be very difficult for DISCOVERY to make the data available through an easy to use set of national standard APIs and query languages. The bar would simply be too high for most application providers wanting to use the data.

To lower this bar and make it as easy as possible for new and existing health care application developers to access and interpret health care information accurately, data should be stored in a common nationally recognised format (for example FHIR or openEHR), with coded items using a standard taxonomy i.e. Snomed-CT.

In order to translate the original primary data into these recognised standard formats, the data has to go through a series of transformation processes. Data transformation is a well-recognised process where data attributes and elements are mapped from one known format to another.

DISCOVERY provides a transformation service and does so as a library of ‘plug-in’ components. Each one of these components provides a different type of data translation, moving data elements from one structure into another. No data loss occurs during this process, and the semantics of the data remain unaltered. Recognised codes are sometimes mapped into common schemes such as SNOMED CT, however no original data or codes are removed; they are simply augmented into the new structure as new elements and attributes.

As there are many systems, and many message formats, DISCOVERY offers an open-source authoring and source code versioning environment whereby many systems developers from different organisations are encouraged to add new and updated message transformations to the overall library. This promotes the accuracy of message transformation routines, with many system experts making valued contributions.

4.4.6 Secondary data store

The secondary data store is the fundamental component within the overall DISCOVERY service. Earlier in this document, the benefits and concepts of the data store are highlighted.

The secondary data store (secondary to the primary clinical data stores provided by the care management systems) is actually the primary data store within the service, used to create additional resources for the various uses.

Section 4.4.2 also explained the benefits of the NoSQL data store technology with regards to high-availability and ultra-scalability.

This section explains the structure of the data store.

The DISCOVERY databases must store data in structured formats. The data must be stored conforming to industry-recognised resource based formats (i.e. openEHR or FHIR).

The structured formats, from a clinical record perspective, are keyed by patient NHS number.

Patient data is stored according to the organisation that controls (owns) the data, but are also partitioned (or to use the technical word 'sharded') by patient ID. This ensures that patient records are horizontally scalable and well balanced across the many data nodes in the data cluster.

The design must support optimised approach to access for both of the following:

a) Organisationally oriented approaches whereby only patient data relating to an organisation (or service) is used
b) Patient oriented, whereby patient data related to multiple organisations can be used subject to data sharing agreements.

Every structured record in the data store contains the core set of attributes as dictated by the primary system, as well as a series of structured 'meta-data' elements which are used by the DISCOVERY system to further filter records based on more complex query and record extraction criteria.

4.4.7 Additional data stores

As well as storing clinical and workflow data in NoSQL data stores, DISCOVERY also stores data / or makes it available for storage in other database formats that may better suit different types of applications.

For example, some analytics languages and applications have a preferred method of data structure best suited to these types of applications.

Other applications may wish to query data in industry-recognised formats, such as relational database management systems.

DISCOVERY provides both APIs and bulk record transmission protocols to allow further databases to store primary system data in other formats for other purposes, as dictated by the data controllers.

Where appropriate DISCOVERY also provides access for applications to these alternate format data stores as part of it's core data service. (See section 4.6)

4.4.8 Data sharing agreements / protocols

Data sharing agreements underpin the data-sharing model of the DISCOVERY data service whenever data is used across different data controllers.

Data sharing agreements are configured within the service in a way that is referred to as data sharing protocols i.e. the service operates according to a technical protocol, which is a set of rules, those rules conforming to a documented sharing agreement which data controllers understand.

It is anticipated that the current “many to many” set of data sharing agreements will be consolidated locally to single inclusive agreements or “Frameworks” to which data controllers consent to using.

Each data sharing agreement or protocol is a set of defined rules that dictate how data can and must be used by systems and applications outside of the primary source systems.

The service never processes or transmits data to health care applications that do not conform to known sharing rules.

The sharing engine uses a series of authored rules to determine the legitimate usage of data as it flows into and out of the data service. A data provider and a data consumer must be listed within a sharing rule before data can be exchanged in any way.

4.4.9 Service monitoring and deployment

Running a large-scale data service requires a 24/7 support services infrastructure.

DISCOVERY provides all the application level monitoring software and user interfaces for assessing and tracing the health of all the components running within the services.

Live dashboards and alerting mechanisms must be in place for personnel to monitor the overall system. Dashboards have metrics for viewing how well the system is running with regards to performance levels, resource usage, warnings and error statuses.

24/7 error and warning thresholds must trigger automatic alerts for rectification action.

Support personnel will be able to use the dashboards and alerting UIs to drill-down into message queues, audit logs, data stores etc. for the purposes of diagnosis and debugging.

All deployment of software components must be strictly version controlled with tight testing and release scheduling.

The software components are backwardly compatible across previous versions to provide the most compatible and stable environment.

The service must be capable of running multiple versions of software components and message formats in parallel.

New and upgraded software must be tested within a separate test environment, and software will be released in phases, using automated scheduled deployment tools.

Bug tracking will be recorded in industry standard tracking tools, and support personnel will have full access to diagnostic reports and debugging environments.

4.5 Standard APIs for consumer applications

As an application agnostic service using standards, the DISCOVERY service can move quickly to provide and publish many APIs as demand determines and as emerging standards are forthcoming.

The APIs in the service are focused on use cases that do not require real time exchange. This includes the many data related use cases which can safely operate with some delay, but do not include the use cases that require real time interactions with the primary system.

APIs must be made available through industry standard exchange mechanisms and protocols such as REST and SOAP.

API standard methods must be aligned with nationally defined and controlled by the NHS.

4.6 Data export and support for analysis applications

As a data access service, the primary purpose of DISCOVERY is to support access supports two approaches to data access for analysis:

1. Export of data to other systems, particularly those that require data to be stored in more specialised formats.
2. Provision of data stores within the service in formats suitable for query by applications selected by the data controllers.

A configuration tool is required to enable these two approaches to be used in practice.

The configuration tool is used to define the parameters for data export or database provision, the configurations operating within the confines of the data sharing agreements that the data requestors are part of. Parameters include:

1. Population definitions

These are the definitions of the population of patients, or workflow items of interest for the analysis. Examples of these are:

a) Patients registered with, or on the caseload of, one or more provider organisations in a geographical area

b) Patients residing in geographical areas

c) Patients with particular code clusters in their records (e.g. diabetics)

d) Workflow items associated with particular organisations

2. Data item subset definitions

This is the subset of data items that will be used in the analysis, the subset being further restricted by the sharing agreements. For example, an analysis of diabetic outcomes may be limited to a data set related to diabetes. There may be no need to include data items not related to diabetes.

3. Data Schema

The type of database schema that the data will be exported or created in.

5 Organisational structures

5.1 Implications of open source and communities

For a DISCOVERY service to be truly vendor neutral it obviously needs to be perceived as being owned and controlled by organisations that are not perceived or categorised as vendors.

Historically, the main NHS vendor neutral organisation has been the HSCIC and its predecessors, Connecting for Health, and NHSIA. There is room for others, working in partnership with the current NHS vendor neutral organisations.

Until recently the dependency on commercial IP, for software delivery has been absolute. However, in the last 10 years the presence of increasingly highly sophisticated and robust open source solutions has resulted in the design and building of solutions in open source across many industrial sectors. In the area of data integration, the technologies have been understood for some time and whilst they have not been brought into the NHS as yet, there is no reason why this should not be the case.

The effect of this is to push the commercial IP boundary towards innovative software and away from legacy software in the same way as new drugs have an initial proprietary phase followed by a generic phase. For data integration software, we are in the generic phase.

Until recently there was also a requirement for commercial organisations to operate the services. However, the NHS has brought more and more service provision in house. For example, spine 2 is not only based on open source technologies but the support of it has been brought into the NHS from the commercial sector, resulting in massive savings.

It should be emphasised that for the vast majority of complex operational systems, whether open source based or not, there is a need for specialist vendors and, as investment risk applies, they are likely to be commercial for the foreseeable future.

The creation of the open source, Code 4 health program and the communities that now exist, has created a collaborative environment that enables sharing of expertise (across and outside the NHS) for design and development without recourse to the commercial IP model.

Although a large system, DISCOVERY is a very small and discrete set of software components tasked only with providing data and not providing advanced healthcare business logic, the responsibility for which remains with the application providers. This means that the new open source approach can be used throughout without recourse to conventional commercial interests. Costs are not eliminated by this route, but are massively reduced over the period of a project.

It is also now practical for the software elements of the service to be managed by the NHS and in particular, by the care providers at the coal face collaborating to share resources.

This pattern has been evolving for some years. The presence of local informatics services has started to demonstrate the benefit of local collaboration, although they remain in general dependent on commercial suppliers for data integration. This dependency is about to change.

Data centres remain in the domain of vendors, but even here we are seeing a massive shift. Community interest companies now run hosted services and the costs are beginning to match. Cloud hosting costs less than 10% of current hosting costs.

Thus it is now entirely feasible to expect that DISCOVERY could be produced without commercial level financial or risk to the NHS.

This document proposes exactly that.

5.2 Implications of emerging standards

Every successful national scale interoperability initiative in the NHS has used open standards. Conversely, there has been no successful nationally scaled interoperability initiatives in the NHS based on proprietary approaches.

However, Open standards have historically been perceived as almost glacial in development rate and often difficult to use and very difficult to understand. Frequently they have not been fit for purpose.

The emerging standards have arrived to challenge this view. With modern approaches to testing and example demonstration, it has brought forward the potential to rapidly develop fit for purpose, easy to use standards. In one emerging standard, the standards authors plan to include the test harnesses in the standard itself.

This lowers the bar for entrants and new developments and in particular allows application development to proceed with much more confidence and as a result, more rapidly.

Open standards creates the potential for eliminating design and development risk, by making sure that alternative approaches can be used in parallel, with confidence that the competitive solutions that emerge will be compatible.

It is expected that DISCOVERY service would be a series of developments creating a number of options at any one time and leap frogging each other in terms of enhancements. The component design of DISCOVERY means that local initiatives can and should be fostered and can proceed independently with backward compatibility being maintained with the central core.

5.3 Organisations

The model envisaged is that of local control and ownership but with pooled expertise and resource in order for it to operate at scale in an affordable manner.

A number of types of organisation are envisaged. The roles are:

1. Custodianship of the technology. Making sure that the IP remains available and that the rules of open source and open standards are followed
2. National service provider. Holds the hosting and core service contracts and is responsible for the hosting services, support of the core technologies, core service levels and core development.
3. Local service provider. Holds the local contracts and is responsible for ensuring that local demands are met, together with support and development of local extensions.

It is envisaged that this is made up as follows and all clinically led:

1. Custodianship. A trusted CIC
2. National provider. A newly form CIC with local providers, NHS England and HSCIC as members
3. Local providers. Various models could include

a) CICs whose members are mad of providers such as practices or federations, trusts and CCGs
b) Consortia of provider organisations
c) Informatics services, or selected CSUs.

6 In Summary

There is a need to do something differently to address the changing requirements of Health and Social Care providers who have a duty to share information.

Most current specialist extractors only extract data sufficient for specific purposes. This limits the benefits, requires a constant refresh of sharing agreements and impacts on the role of the data controller.

The Vendor Neutral Data Access Service offers a solution to provide access (by the data controllers) to an up to date logical equivalent of patient and related data held in their care management systems.

The service shall be owned and managed by vendor neutral service providers, including the healthcare providers themselves.