

Technical Improvements and future directions for eCollection and multimodal data capture in the Australian Bureau of Statistics (ABS)

Author and presenter: Monica Kempster, Australian Bureau of Statistics

1. Abstract

The ABS is positioning itself strongly for increased demand for electronic collection of data, which will support the Australian Government Digital Services Strategy.¹ The ABS is further improving eCollection capability, utilization and usability for data providers. This includes adopting a self-help user model, user self-management of accounts and credentials, improvements in multi-modal data capture and management.

Significantly more surveys have been added to the electronic capture program; with many requiring new layouts to improve statistical collection, reduce modal bias, improve accessibility compliance and provide value add functions such as address validation, through the utilization of corporate geospatial services. Currently the ABS has been migrating forms to the eCollection environment under a 'Translation Phase' which means that the layout and questions are created as a direct copy of what is currently sent out under our paper or personal interviewer modes. As the ABS moves into a transformation phase for its acquisition of data, the ABS will take opportunities to utilize the new eCollection mode to enhance data collection through the use of web Services, layout questions to be more usable in a browser or device format, and potentially provide users with the opportunity to provide more data than what was previously collected due to paper form constraints. The use of Blaise in the transformation space for data collection, editing and analysis is instrumental to the success of this phase of the project. Through the work done over the last 12 months, an increased focus on multi-modal data capture and processing has become a core part of development and improvement.

The ABS is also adopting a new corporate authentication and authorisation solution. Changes have been made to the existing Blaise solution to improve connectivity with other components, such as authentication, security and processing services. There are also improvements in system security, load and performance. This work will make the Blaise instrument development and linkage more robust for the future, through the exclusion of all non-form processing services to servers outside of the Blaise server park environment and making the instrument more resilient to external ICT changes.

Future improvements include the ability to move partially completed records between data collection modes, making multi modal capture more timely and adaptable to changing data provider behaviors, as well as enabling scale-out of infrastructure to multiple Blaise parks for load balancing, risk management and contingency planning.

2. Background

The ABS has made a strategic decision to use Blaise IS for all web form development, as well as continuing its use for the interviewer collected household collections. Current web form development is using Blaise 4.8.4 (upgrading to latest builds approximately every 9 – 12 months). The ABS released the first electronic form (eForm) into production in December 2012, and has continued a migration work schedule since then, moving both business and household forms into the environment.

Across the last two years a number of requirements have come up for adding additional functionality as more complex forms are migrated. These have included requests for more complex layouts (to support matrix style responses) and enrichment of feature sets through the use of external services. External services such as the ABS Address Validation service, which has been linked into the form, to provide address validation in Agriculture forms. This will improve the ABS ability to utilize geospatially enabled data in publications and micro level datasets.

The ABS has also been developing a corporate Authentication and Authorisation solution to be utilized across all externally facing systems. The External Identity Access Management (XIAM) solution will be first utilized in the eCollection space. It will move users to a system of self-managed credentials, as well as providing features to update their contact information. The XIAM system will also provide obligation management functions such as delegation of forms between registered users and a dashboard to show total obligations and their statuses. XIAM will also deliver the ability to group a number of form obligations into a single set for management, which will improve the management of data providers with large numbers of forms and significantly improve the migration of surveys that use sets of forms to collect information.

The XIAM solution will decouple Blaise from authentication and authorization services. This will provide the ability to utilize multiple Blaise parks to produce a technical solution which enables the ABS to:

- scale out performance
- reduce complexity
- manage surveys
- exclude critical or high volume surveys onto a separate Blaise park, and
- manage external components such as provider portals in a more modular way.

External factors such as compliance with Accessibility Guidelines⁴ which is expected of all Australian Government agencies are also generating changes in the delivery of eForms. The ABS has made some improvements to both our code and the 'out of the box' CSS, however, there is still more to do be done to meet the expected level of compliance.

3. Current Architecture

The following diagram shows the current architectural solution overview for the Blaise eCollect platform. The diagram depicts all major solution components as well as major communication flows.

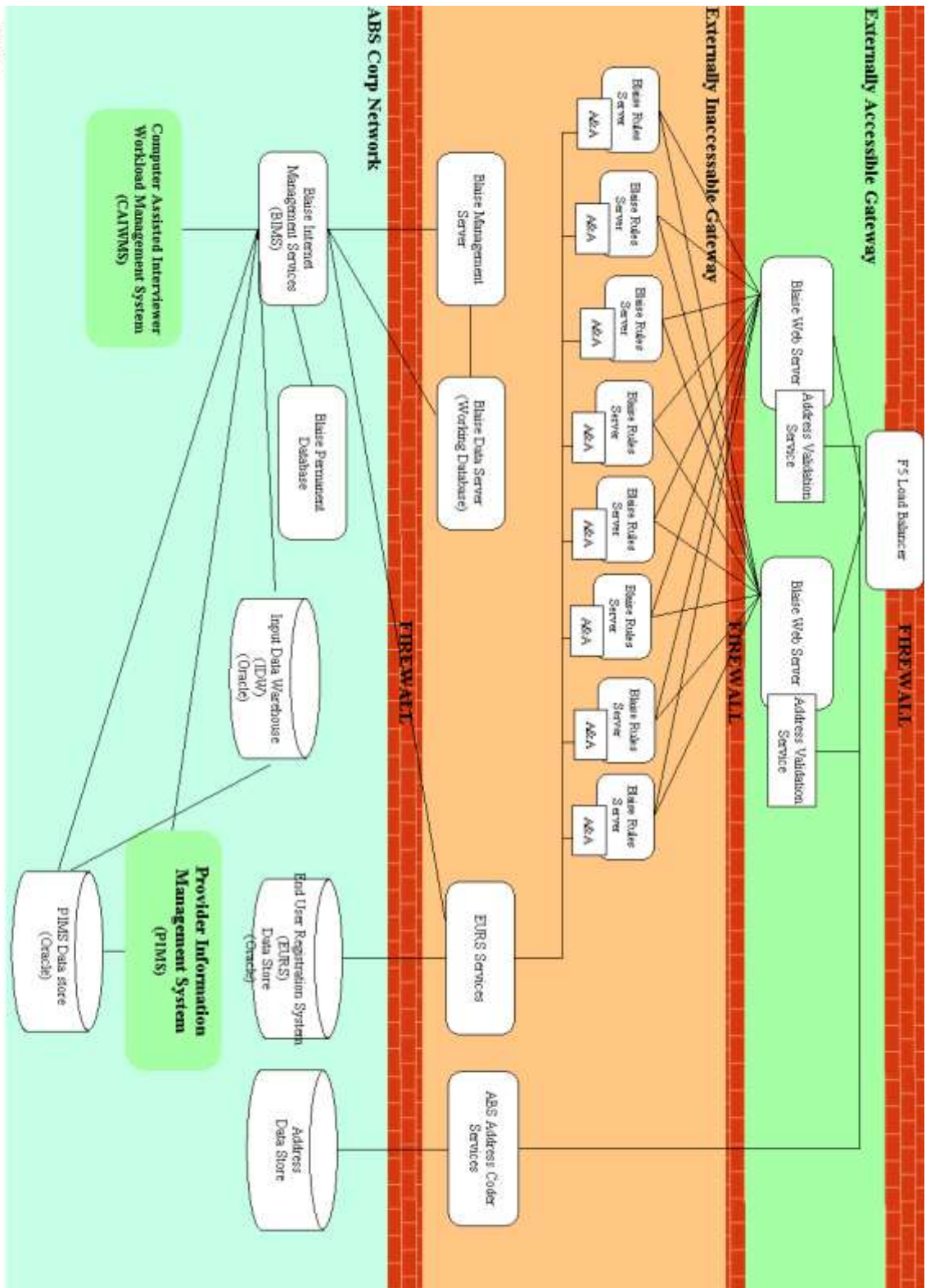


Figure 1 Current Blaise eCollection Solution Architecture

3.1 Components

The ABS ECollection solution utilizes the ability to put the Blaise IS components onto multiple servers to improve performance by utilizing multiple server capacity. The number of scaled out servers for each component has been determined based on load testing, and optimized to maximize the throughput and stability of the system.

Notes:
 A&A Refers to the Authentication and Authorisation DIH
 Links of the Blaise Data Server other Blaise Servers (Web & Rules) have been removed to improve readability of diagram
 Links of the Blaise Management server to all other Blaise Servers (Web, Rules, Data) have been removed to improve readability of diagram

The current environment configuration of 2 Web Servers, 8 Rules Servers and one Data server is what ABS found to be its most optimal configuration through load testing, for more details see the 2013 IBUC Paper, Blaise 4.8 Load and Performance Testing². This architecture maximizes the total number of connections which can be handled by the Data Server (of which there can be only one in a park) while ensuring that no other component experiences performance degradation. As can be seen by the fact that there are 8 Rules servers, this server experiences the highest load of all Blaise servers.

3.1.1 Blaise Web Servers

The Blaise Web Servers provide the presentation layer for Blaise Internet. The Web Servers interact with the Rules servers when a user navigates to the next survey page, when a critical field is updated, or a user submits a survey. We also have the Address Validation Service installed on each of the Webservers.

3.1.2 Blaise Rule Servers

The Blaise Rules Servers are responsible for executing the business logic for web surveys. For example, determination of the next sequence of questions and compute derivations. It also invokes the authentication and authorization DLL calls via alien procedure calls within the instrument.

3.1.3 Blaise Management Server

The Blaise Management server coordinates the Blaise Server Park for load balancing, and provides functionality for deployment and management of surveys deployed to Blaise Internet. In the ABS production environment all Blaise Management Server functions are accessed via the BIMS Services (discussed in more detail in section 3.1.6). The current architecture also means that the Blaise Manager is within the gateway environment, and not accessible to anyone other than a few members of the ABS Technical Infrastructure team. By utilizing the BIMS Services (which are in the corporate environment) other users of the Blaise eCollection systems can invoke Blaise Management Server functions from within the corporate environment. This has been done in line with our security policies and architecture principles.

3.1.4 Blaise Data Server (Live Database)

The Blaise Data Server (Live Database) stores collected provider data for the duration of a survey; this includes both partially completed responses, as well as submitted survey data. Completed survey data is then copied (via BIMS) into the Permanent Data store. Access to this database is only via the BIMS Services for the same reasons described in 3.1.3 above.

3.1.5 Blaise Permanent Database (Offline Database)

The Permanent or Offline data store is used to store provider response data once it has been submitted. This component is used to mitigate security risks associated with storing provider data permanently in the live data store (see 3.1.4), due to the fact that it is in the gateway environment. It also reduces load risk; as more intensive data management functions such as load to our Input Data Warehouse (IDW) do not have any impact on the performance of the online systems. This database is also within the corporate environment which makes it more accessible to other systems, and functions.

3.1.6 Blaise Internet Management Services (BIMS)

The Blaise Internet Management Services (BIMS) is an ABS built web service layer that provides interfaces to survey lifecycle management functions in the Blaise API, and links to PIMS, CAIWMS and Authorisation functions.

The Blaise Internet Management Services provide a single point of access for other system to access Blaise Functions. This has been done to provide a ‘wrapper service around Blaise’ to enable the Blaise API calls to be obscured from other systems, and provide a consistent interface access for other systems, and limit changes which may be required to these services when Blaise upgrades occur to a single place within the environment.

Operations that are facilitated by BIMS are deployment include

- deployment, update and decommission of surveys
- update of prefill data
- retrieval of processing statuses, and
- access to provider response data and Blaise journal functions.
-

BIMS interacts with other ABS systems such as

- the Input Data Warehouse (IDW),
- Provider Information Management Systems (PIMS),
- External User Registration System (EURS) and
- Computer Assisted Interviewer Workload Management System (CAIWMS).

The PIMS and CAIWMS systems provider the user interfaces for people to invoke the BIMS Services such as initialize survey, prefill data retrieve data and access journal data.

3.1.7 External User Registration System (EURS)

The External User Registration System (EURS) is an ABS bespoke developed system, which uses an Oracle database and JBOSS developed services to provide Authentication and Authorisation functions. This system is used for the ABS eCollections and dissemination products. EURS is currently utilized in the eCollection environment to provide a pre created user identifier and password to eCollection data providers, and already linked survey obligations.

3.1.8 Authentication and Authorisation Module

The Authentication and Authorisation Module is an ABS developed component. Its purpose is to provide the link between Blaise and the EURS Services, given that Alien Procedures will only call .Net, it effectively provides some business logic and a wrapper for the Java based services. It is installed on each instance of the Blaise Rule Servers. It is used for Authorisation services, as well as to write back the completion status of a form to the EURS data store

3.1.9 Provider Information Management System (PIMS)

The Provider Information Management System (PIMS) is used to manage business surveys within the ABS. It contains the selected units, response statuses, workflow management for Intensive Follow Up (IFU) and sample management. It is also used to initiate and document provider contact information for the management of surveys. It is used in conjunction with the Input Data Warehouse (IDW) which stores the provider responses. There are other internal systems which are used to manage this further, but will not be discussed in this paper.

3.1.10 Computer Assisted Interviewer Workload Management System (CAIWMS)

The Computer Assisted Interviewer Workload Management System (CAIWMS) is the ABS system to manage Household Surveys. It contains the interviewer workload allocations, response statuses, links to coding systems and the respondent data, which is cleaned, coded and then extracted. This is where the survey initialisation for household forms is done, through the CAIWMS interface, linking with

BIMS to create the user registrations in EURS, and copy prefill data into the online data stores for reference when someone accesses their forms.

3.1.11 Input Data Warehouse (IDW)

The Input Data Warehouse is an ABS developed solution, built on Oracle to hold micro data for all business surveys. This data store is loaded with the eCollection information for business surveys via a BIMS call to the IDW loader.

3.2 Current Challenges and Limitations

Discussed below are some of our key challenges in the architecture and system environment, which we are currently addressing.

3.2.1 Authentication and Authorisation

Briefly, Authentication is providing a way for someone to prove who they say they are, usually through the provision of a user identifier and password matching combination. Authorisation is what an authenticated user has access to within the environment.

The eCollect authentication and authorization is currently a bespoke system, which uses EURS to provide Authentication services, and an Authorisation system developed in Blaise to persistently store form obligations, and to provide the user interface for both Authentication and Authorisation functions.

There were a lot of additional checks put in place within Blaise to ensure that the security of the system was maintained, and things like session and URL tampering could not result in a security breach. While this solution has served for the last two years in enabling the ABS to initially move to eForms, the desire for a more featured user portal and whole of organization approach to External Registration and Authorization has led to the development of a new system.

Reasons for this are discussed briefly below.

- The current method of providing businesses or households with a user identifier and pre generated password means that the ABS does not necessarily have a link to the person filling out the form, only the address of a household, or the primary contact at a business (who may not be the data respondent)
- Providing a user with an identifier and password means that if the password is forgotten, or the account is locked, there is no meaningful way to provide a challenge/response check (i.e. secret question and answer, or email token) to provide a secure way of identifying the authorized requester to reset a password
- Use of already existing components in industry is now the preferred approach to the creation of new ICT systems within the ABS, reducing our dependency on bespoke systems, such as our internally developed Blaise solution
- For large ‘fleets’ of responses, such as Survey of Motor Vehicle use, a data respondent could have 30-40 motor vehicles to respond for. At present this has not been moved onto eCollection, as the current method of providing 30 ids with 30 passwords (one for each obligation) is not seen as an appropriate end user experience
- The ability to provide a ‘portal’ experience to users, where they can see all of their obligations, their due dates and status, is expected to improve users ability to manage their responses in a timely, consistent manner
- The use of a separate authentication and authorization solution will improve security of the system, as newer technologies can be used to ensure that URL tampering and session tampering cannot be used

- Moving authentication and authorization services away from Blaise will reduce the load on the Blaise servers, improving load performance of the Blaise services.
- People can create a persistent account to re-use credentials for additional survey obligations – including a way to link both business and household obligations into a single
- The ability to delegate an obligation to another registered user, currently the only way that can be done is by sharing credentials, which creates a security risk, and obscures management information which would assist the ABS in understating user behaviors when completing survey obligations
- Use of a single solution across the ABS for Authentication and Authorisation means when security requirements change (e.g. password lengths, or complexity) a single system can be changed)
- The solution should be more scalable than the previous solution
- The removal of authorisation from the Blaise environment opens up opportunities to have multiple Blaise parks, as the authorisation information will be in a data store which isn't within the park.

3.2.2 Data Retrieval of only Completed forms

Currently systems are only designed to retrieve submitted data, as our legacy systems were not updated to handle the retrieval of Web forms which are partially complete, and moving them into an interviewers' workload, this opportunity has come about through the utilization of multi modal data capture. While services could be built to retrieve any data, challenges in legacy systems have made this a future enhancement, rather than a current functionality.

The business challenge to manage while this technical limitation exists is that respondents, who may have significantly completed a form online, will be re-asked all of that content again by an interviewer, possibly increasing respondent frustration, and not maximizing the use of interviewers to only complete the data which hasn't been already collected.

3.2.3 'Plug and Play' Address Validation

The current implementation of address validation was an initial technical delivery. At present it is tightly coupled to the name and position of the Blaise field within the instrument, which means that each instrument which wants to implement this service either has to put it in the exact same named field, with the same designed page, or we need to add those field names into the Validation Services. Work is being done to decouple the exact Blaise name from the call and create a plug and play module, so that it can be packaged into the Blaise Instrument and use a generic service, which can be referenced on a separate server, rather than depending on it being deployed on the Blaise Web Servers to enable communication

4. Proposed New Architecture

The architecture discussed below is currently being developed, with most of it in user acceptance testing. It is expected that this will be the production system within the next 6 months

4.1 Components

The discussion below focuses only on the improvements made to the system. Given that it is expected in the short to medium term for the ABS eCollection to run on Blaise 4.8.4 the dimensionality of the Blaise server park has not changed. The details of some components have been removed from this section as there have not been any changes made.

4.1.1 Blaise Web Servers

The Blaise Web Servers will be set up the same as the current architecture; however, the Address Validation Services will be moved onto a separate server. This will move the load which is currently being managed by the Blaise Web Servers onto a separate server, which means that non eForm use of those services won't have any impact on the Blaise Server performance. It will also make deployment of newer versions of the Address Validation Services easier, as it won't require taking down part of the eCollection environment to update the service.

4.1.2 Blaise Rule Servers

The Blaise Rules Servers will have the Authentication and Authorisation DLL removed, and it will be replaced with a Web Services which will be called from ASP.

In the short term for Initial production release of XIAM, an OMSServicesUser.DLL will be put in place this will write survey completion statuses to the OMS store. It will be deployed in the same manner that the current Authentication and Authorisation DLL is architected.

4.1.3 Address Validation Services

The address validation services are effectively gateway enabled wrappers to enable linkage with internal ABS Services, to enable them to be consumed by Blaise within the gateway. These have a potential future user base much larger than just eCollection, including geospatially enabled data, Census data searches via address, and National Regional Profiles. These address services will be rewritten to de-couple the Blaise names and field indexes, to enable all Blaise Instruments to use a plug and play approach to calling these services. The movement of these services onto a separate server in the gateway will also make them more usable broadly in the ABS.

4.1.4 Obligation Management Services (OMS)

The Obligation Management Services (OMS) is replacing some EURS components of the old system which will store the obligations that a user has. Storage of Authentication and Authorisation data has been moved into separate data stores for the new release, allowing authentication to be managed as a Corporate System, and authorization (Obligation Management) to be built for each application depending on unique requirements.

OMS uses an oracle database which is configured in the corporate domain, and provides the ability for the acquire portal to make changes to obligation records in the data store via this service.

4.1.5 PWM

PWM is an open source password self-service application³; it uses an LDAP data store and provides a basic set of customizable password management features. This has been chosen as a 'buy before build' option to move the ABS into our future method of building systems. It provides the setting of secret question and answers for credential management, reset of password, and forgotten password services.

4.1.6 Access Policy Manager (APM)

Access Policy Manager or APM is a F5 device plug in which provides the ability to manage authentication (i.e. credentials to allow access) this provides the create account and login functions. This solution promises to be highly scalable, available and be able to be reused across the ABS.

4.1.7 Acquire Provider Portal

The acquire provider portal has been built to provide the interface layer where users can link an obligation, access a dashboard of their obligations, update their email address and call change password functions. They can also delegate obligations to other authorized users.

From the obligation dashboard a user can then click on the survey link and access the Blaise Instrument in order to complete their response and submit their form.

4.1.8 Multiple Server Parks

The removal of the Authorisation being in a Blaise Instrument will mean that with some changes to BIMS deployment metadata (to enable it to capture which park an instrument is deployed in) the eCollection system will be capable of managing multiple Blaise parks. This will be useful as a means to be able to handle additional load, by increasing the capacity of the eCollection Blaise Instrument component, as well as provide the ability to be able to exclude particular surveys onto a separate park (for example the Agricultural Census scheduled to be delivered in 2016). At this point there is nothing which suggests that extension beyond 2 parks is not possible, however load limits on the BIMS services would need to be tested and consideration given to a load balanced pair of BIMS Services.

5. Brief overview of Blaise Instrument architecture

The way instruments are architected allows for information which is already known about the form, such as the selection and historical information. This information is known as prefill data to be utilized within the form. For a diagram depicting the relationship between Blaise source code and compiled packages see Figure 3

The BIMS services are used to load the prefill data from either the PIMS environment (Business Surveys) or CAIWMS (Household Surveys) and it is written into the Working database server, as a _INIT named database (for each survey).

For some surveys data is loaded from CSV files into databases which are packaged into the instrument. This is done for surveys where we have information from other bespoke systems, which is required to provide data respondents context of what they are to report on, for example building jobs for the Building and Construction Survey.

The instrument package is deployed via BIMS, the internet BIP file is used to deploy the instrument out into the Blaise server park, and the other version of the code is used to create the metadata for CAIWMS. BOI files are also created to enable BIMS to access the INIT, Journal, Permanent, Temporary and Work data stores

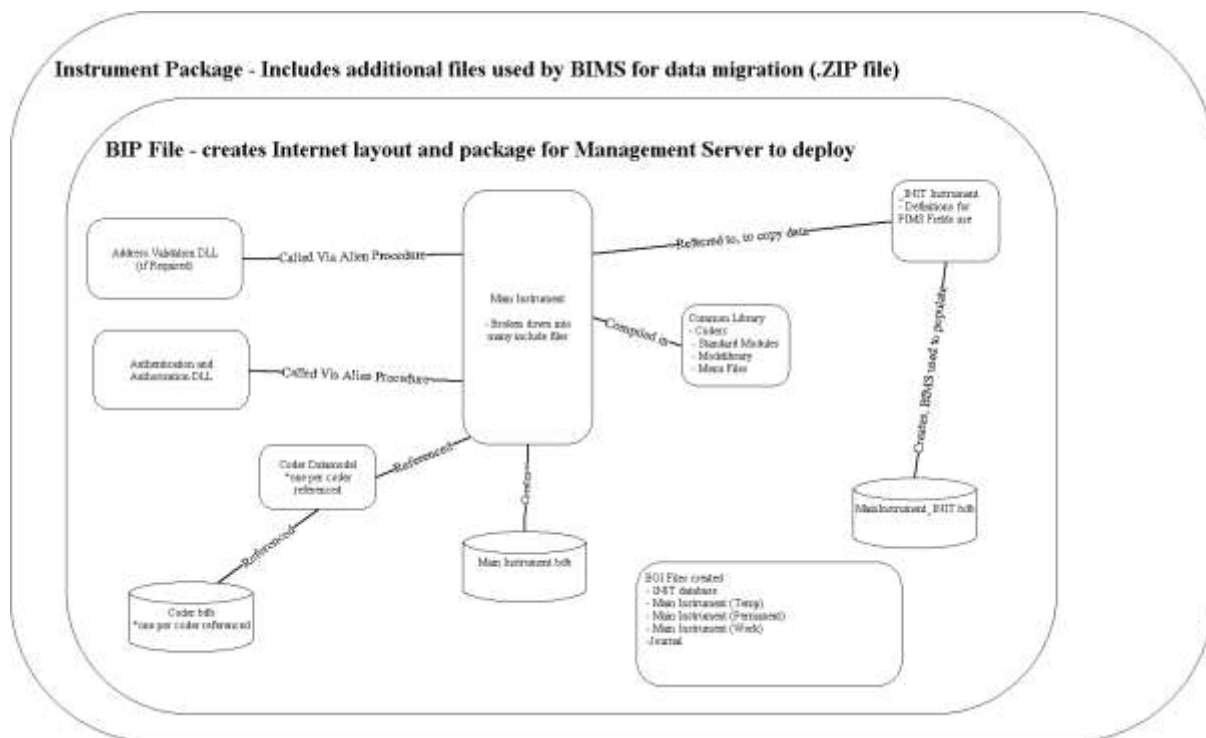


Figure 3 – Blaise Instrument Architecture

6. Accessibility Compliance - Issues, Challenges and Improvements made

There are still a number of challenges which are not directly to do with the large scale architecture. As discussed before, all Government online systems are expected to conform to WCAG 2.0 AA Standards. While the ABS has made some changes to bring us closer to conforming with these standards there is still more work to be completed.

There has been a tension between usability and accessibility. Some of the decisions made to improve the layout and usability of the form have had a negative impact on the ability of the form to be read and understood by users of screen readers. For example improved layout has decoupled the question wording from the data entry field, which provides a better looking interface, but makes it more challenging for screen reader users to match the question with the response field.

6.1 Accessibility Compliance Issues and Challenges

Some issues which have been raised through Accessibility Testing which we have currently not solved are:

- Menus, Headers and Footers – they are not marked up as lists, so are not able to be easily navigated
- Resize of text using Browser zoom
- Pages sharing the same title – each page uses the Data model name for the title
- Change of context – occurs when rules are re-run on a page (currently happens on re-derive of field, or on error being triggered)
- Error message not being read by screen readers
- Label_ids are not meaningful but are used by screen readers
- HTML code does not validate appropriately (when checked with W3C Validator)
- Repeated content in headers and footers

- Radio button selection and check box selection is not consistent in whether you can click on the line to select

6.2 Accessibility Compliance Improvements Made

Changes which have been made which have improved accessibility compliance

- The user of the layout=presentation for tables which are implemented for layout purposes only
- Text, rather than text as images on buttons
- Alt text has been added to all images
- Improvement of layout of forms so that screen readers keep context together
- Using Header Markups to enable users to scan through the page to provide a structure which can be quickly navigated
- Colour Contrast
- Role of pages is no longer set to ‘application’

In order to continue to improve our compliance, further training of staff to understand the guidelines, utilize tools for assessing compliance and understanding of how to use tools like JAWS for screen reading will assist.

7. Management Information

An e-Form flag has been added to the Labor Force output data, to enable analysis to be done based on the mode that the data has been collected in. This is a first step toward having more management information, processing statuses and multi modal management available.

This flag has been enabled due to increased divergence between CAWI and CAPI Instruments, in respects to question wording, questions asked of respondents, sequencing of questions, derivations and fields which are on path.

The reason for having an e-form flag available in the environment is to:

- provide the ability to analyse the data, and determine if there is any modal bias in responses to the instrument
- Manage which edits within the instrument are triggered for eforms when they are bought into the processing environment (i.e. don’t run CAPI only rules which aren’t triggered in the CAWI Instrument)
- Provide abilities to select different data items, or processes based on the mode of collection

We are currently only in early stages of implementing this, given that the CAIWMS is a legacy system, and changes can be difficult to manage effectively. It is also important to ensure that changes are not going to adversely impact the processing of data.

8. Alerting, Monitoring and server restarts

There have been instances where the ABS has experienced instability in its systems. Due to this there are now regular restarts of all servers, to ensure any memory leaks, or threading issues are regularly reset in the environment.

Some of these restarts have been put in place as a temporary measure to enable the ABS to thoroughly investigate the issues, and be assured that the fix is operational before making any changes to the production environment.

Automated test scenarios and alerts have also been set up on the servers, and a number of ABS staff across Technical Applications, Technical Infrastructure, Household Instruments and Business

Instruments are now alerted should an outage occur so a rapid response can be taken, and impacts and actions can be managed.

9. Modular Design

The desire of having ‘plug and play’ components within the ABS is also extended to Blaise instruments. The ability to quickly deliver instruments to collect various different topics, and assemble already created components, rather than delivering an instrument from a new specification will assist with driving the ABS forward with the goal of fast turnaround on statistics improving time to field, use of already tested questions and sequencing and known working components.

This is driving forward the requirement to build instruments in blocks and assemble, rather than having large, monolithic Blaise instruments, which are built for a single purpose.

Having components, such as the Address Validation Tool, as plug and play modules will also further this capability, and make linkages to rich services into instruments a more straight forward approach, and the technical complexity obscured within the black box component.

As part of the Acquire transformation phase the ability to manage obligations which have a dependency order for completion and provide seamless transition between these components is being investigated. An example of this would be the Monthly Population Survey which for a respondent in a given month may contain both the Labor Force survey, as well as a supplementary survey, both of which are required to be answered, with the Labor Force needing to be completed before moving onto the supplementary, as data is transferred and referenced in the supplementary from the Labor Force.

10. Summary

Significant work has been done to fine tune the eCollection infrastructure to improve scalability, performance and reuse. The current work to improve the Authentication and Authorisation solutions will provide a number of benefits such as improved user experience, encourage self-management of accounts, and obligation management, and also provide significant opportunities to further scale out the eCollection architecture to continue the migration of ABS Instruments into online data capture.

The ABS will continue to work towards improving Accessibility compliance, both through Blaise Coding improvements, CSS templates and looking toward future Blaise releases to assist in providing opportunities to make changes to improve conformance.

Future work to improve modularity of Instruments is progressing to allow the ABS to adapt and meet new demands in a rapidly changing environment.

11. References

1. Australian Government Digital Strategy, Accessed 16/03/2015, <https://agds.test.govspace.gov.au/dss/>
2. Oleg Volguine, Blaise 4.8 Load and Performance Testing, 2013, Accessed 22/03/2015, <http://www.blaiseusers.org/2013/papers/2013Proceedings.pdf>
3. PWM, Accessed 22/03/2015, <http://code.google.com/p/pwm/>
4. WCAG 2, Accessed 22/03/2015, <http://www.w3.org/TR/WCAG20/>