Using Live Data for Testing purposes

# Using Live Data for Testing purposes

## Summary

This document describes the use of live data during testing of Ministry of Justice (MoJ) systems. In general, using live data for testing purposes is considered bad practice. By default, the MoJ does not permit testing using live data. It is highly likely that simply using live data for testing purposes would not be compliant with GDPR.

Following this guidance will help you avoid problems, but cannot guarantee that you have addressed all the concerns. You must carry out a full Data Protection Impact Assessment.

## Who is this for?

This guide is aimed at two audiences:

1. The in-house MoJ Digital and Technology staff who are responsible for testing systems as part of technical design, development, system integration and operation.
2. Any other MoJ business group, agency, contractor, IT supplier and partner who in any way designs, develops or supplies services (including processing, transmitting and storing data) for, or on behalf of the MoJ.

## Do you really need to use live data?

According to [Information Commissioners Office](https://ico.org.uk/), you may use either live or dummy data to test your products so long as they are compliant with data protection law. However, using dummy data may be preferable as it does not carry any risk to data subjects.

If you are processing live data, you will need to complete a Data Protection Impact Assessment beforehand if there is a possibility of risk to the data subject. The ICO has helpful information about using a [Sandbox](https://ico.org.uk/for-organisations/the-guide-to-the-sandbox/) to help utilise personal data safely.

Data used for testing purposes must have characteristics that are as close as possible to operational data. But that is not the same thing as needing to use live data.

Check whether you really need to use live data, by considering the following questions:

1. **Speed:** What are your time requirements for test data provisioning?
2. **Cost:** What is an acceptable cost to create, manage and archive test data?
3. **Quality:** What are the important factors to consider related to test data quality?
4. **Security:** What are the privacy implications of these two sources of test data?
5. **Simplicity:** Is it easy for testers to get the data they need for their tests?
6. **Versatility:** Can the test data be used by any testing tool or technology?

The best test data simulates live operations data.

Note: It is important that test data is protected to the same standard as the live data. This is to ensure that details of the system design and operation are not compromised.

To protect test data, the following principles should be followed:

* The test manager must authorise the use of test data.
* Test data should be erased from a testing environment immediately after the testing is complete or when no longer required.
* The copying and use of test data should be logged to provide an audit trail.

Note: In the absence of an allocated test manager for a project, refer to the system owner.

By default:

* Data used for testing must not contain any live data.
* Using live data containing personal information is prohibited.

In exceptional circumstances, the use of live system data may be permitted. Permission to use live data is by exception only. A valid business case must be approved by the MoJ CISO, system assurer and the Information Asset Owner (IAO).

The Information Asset Owner must ensure that live data will be used lawfully, fairly and in a transparent manner in the interest of the data subject.

A thorough risk assessment, and a Data Protection Impact Assessment, should be carried out to ensure where interdependent applications, systems, services, APIs, BACS, XML, or processes, may be required, these are appropriately reviewed and security controls put in place.

## Anonymising data

It might be acceptable to 'anonymise' the live data such that it can be used more safely for testing purposes. Consider:

* Is it possible to do this?
* What processes can you follow to generate acceptable data?
* Is randomisation sufficient?
* What about obfuscation?
* When is production-like data acceptable (or not) for testing purposes?
* How do you ensure that production-like data is sufficient for testing purposes?
* What are the expectations regarding suppliers - for code, and for services?

If you are considering the anonymisation option, pay particular attention to specific types of data that are often sensitive. Examples of data that must be anonymised include:

* personal data revealing racial or ethnic origin
* personal data revealing political opinions
* personal data revealing religious or philosophical beliefs
* personal data revealing trade union membership
* genetic data
* biometric data (where it can be used for identification purposes)
* data concerning health
* data concerning a person's sex life
* data concerning a person's sexual orientation
* data concerning criminal offences
* email addresses
* bank details
* telephone numbers
* postal or residential addresses

This list is not exhaustive.

In general, recommendations for anonymising data include:

* Replace with synthetic data.
* Suppress (remove) or obfuscate.
* A useful link for anonymising telephone numbers is [here](https://www.ofcom.org.uk/phones-telecoms-and-internet/information-for-industry/numbering/numbers-for-drama).

## Data Privacy considerations

The use of live data for testing, where the data contains personal information, is almost certainly incompatible with the initial specified, explicit and legitimate purpose(s) known to data subjects. In effect, the data subject didn't know that their data would be used for test purposes.

There are sometimes valid reasons when you do need to use live data for test purposes but they are normally the exception rather than the norm and typically looked at on a case by case basis where appropriate risk management calls can be taken.

Looking at datasets being pulled out of databases are a prime example of where you may need to use live data to make sure that a software application is functioning correctly. For some things it is not always possible to use synthetic data.

Where a project is considering the use of live data for test purposes, it is essential to understand the data first, to be clear about what GDPR related factors apply.

You might need to look at fair processing notices and take these into account around the context of the tests being performed.

Note: It may actually be illegal to perform planned tests if fair processing notices do not allow using the data for test purposes.

Where the data involves personal information, help must be obtained from the MoJ Data Privacy team. At the very least, you must revisit or update an existing Data Protection Impact Assessment.

If there is no option apart from using live data, some of the things that should be considered will include the following:

* How will the data be extracted or obtained, and who will perform or oversee the extraction? What clearance do they have?
* What controls are in place to extract the data?
* Where is the data going to be extracted to? In other words, what media or mechanism will be used? For example, is the data extracted using electronic means such as SFTP, or is the data extracted to removable media, or does it remain 'in situ'?
* How is this data going to be protected at rest and during transit?
* What systems will the data be copied to, and in what environments?
* What systems will the data be processed by?
* How will access to this information be controlled both at rest and during transit for all systems that are involved in processing it?
* What access controls are in place end to end?
* Once testing is complete how will the data be removed/destroyed? What assurances do you have over this?

If live data is being used for test purposes within the Production environment, then backups are key and the testing to make sure that backups can be quickly restored is a must. There needs to be a good rollback plan in place. There also has to be an appetite for risk acceptance.

### Ensuring test data is GDPR compliant

If you are intending to seek a special exception for using live data, or if you have anonymised the data but still want to have a satisfactory level of Data Privacy consideration, the follow points will help. Ensure that your test model has:

* Well-defined documentation of personal data information in all test environments.
* Effective data discovery to understand and unearth sensitive data information.
* Implemented a test data management process for the entire data life cycle that includes profiling, sub setting, masking, provisioning and archiving data in test environments.
* An irreversible 'on-the-fly' data masking process for production data within a repository.
* Permission and alerts in place for data exports and access outside the region, as this is restricted.
* Controls to prevent access to personal data from unauthorised access points, devices, or locations.

## If testing is to go ahead

### Developer access

In a normal working environment, developers working on an application, platform or service would be segregated away from access to live/production data. They would never be able to see or manipulate this data. The use of live data for test purposes would potentially negate or bypass these controls.

Also, developer roles are often specified as not requiring [SC clearance or above](minimum-user-clearance-requirements-guide.html). This applies also to external (3rd party) software suppliers generating bespoke applications or services. The expectation is that the developers do not ever have access to live data.

The use of live data for testing may mean that the clearance levels for developers on a given project would need to be reviewed.

### Preparing for tests

Any code or tests involving live data should ensure the following:

* Code performs input validation.
* Output is correctly encoded.
* Full authentication and authorisation is in place.
* Session management is in place to ensure that code and data is not continually available outside the testing activities.
* Strong cryptography is used to protect data 'at rest', 'in transit' and 'in use'.
* All errors and warnings generated by applications, services, or recorded in logs are monitored, captured and actioned.
* A Data Protection Impact Assessment has been performed.
* Any backup processes will correctly filter out or otherwise protect the live data within the test environment.