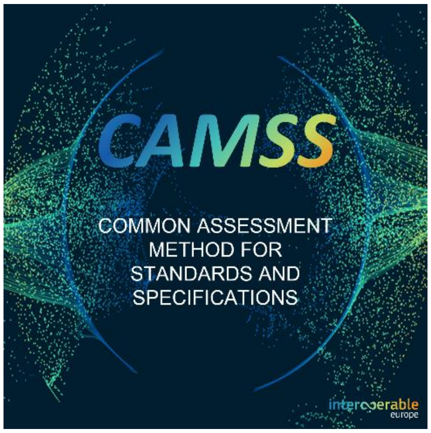

ASSESSMENT SUMMARY v1.0.0

UTF-8[1]

IETF[2]

[1] UTF-8 specification: https://datatracker.ietf.org/doc/html/rfc3629

[2] IETF website: https://www.ietf.org/

Change Control

Modification	Details
Version 1.0.0	
Initial version	

TABLE OF CONTENT

1. Introduction

The present document is a summary of the assessment of **UTF-8** carried out by CAMSS using the CAMSS Assessment EIF scenario[3]. The purpose of this scenario is to assess the compliance of a standard or specification with the European Interoperability Framework (EIF)[4].

2. Assessment Summary

UTF-8 (Unicode Transformation Format, 8-bit) is a character encoding scheme that is widely used for representing characters from almost all scripts and languages in use today. It is part of the Unicode standard, which is a global industry standard for representing text and symbols from all writing systems in a standardized way.

UTF-8 is widely used in web applications, operating systems, and programming languages, and it has become the de facto standard for text encoding on the internet. Its popularity is due to its flexibility, backward compatibility with older encoding schemes, and its ability to support almost all scripts and languages. It was first specified by Ken Thompson and Rob Pike in the Plan 9 operating system in 1992, and later standardized by the Internet Engineering Task Force (IETF) in 1993.

2.1. EIF Interoperability Principles

Interoperability principles are fundamental behavioural aspects that drive interoperability actions. They are relevant to the process of establishing interoperable European public services. They describe the context in which European public services are designed and implemented.

The specification fully supports the principles setting context for EU actions on interoperability:

- **Subsidiarity and proportionality**
  UTF-8 is included in 10 national catalogues of recommended specifications, among which we can find the Netherlands and Sweden. The National Interoperability Framework (NIF) of these Member States is fully aligned with at least 2 out of 3 sections of the European Interoperability Framework (EIF) according to the National Interoperability Framework Observatory (NIFO) factsheets[5].

The specification fully supports the principles setting context for EU actions on interoperability:

- **Openness**
  By ensuring the correct representation and processing of characters from a wide range of scripts, UTF-8 can help to the publication of data as Linked Open Data (LOD). Like all IETF standards, UTF-8 is a free and open technical specification, licensed on a Royalty-free basis. Stakeholdes

[3] CAMSS Assessment EIF Scenario 5.1.0: EUSurvey - Survey (europa.eu)

[4] ISA^2 programme: https://ec.europa.eu/isa2/eif_en

[5] NIFO factsheets: https://joinup.ec.europa.eu/collection/nifo-national-interoperabilityframeworkobservatory/digital-public-administration-factsheets-2022

contrbutions are at the core of Its standard development process, and evolves through an iterative process where draft versions of the specification are made available for public review and feedback comments.

UTF-8 was launched at the beginning of 2000s, and it currently is one of the three encoding methods recognised by Unicode or web languages. This fact demonstrates the maturity and the market acceptance for development of products and services.

- **Transparency**

  By allowing the encoding of data and information through the different information systems and the internet, UTF-8 fosters the visibility and comprehensibility of administration services and eases the decision-making of public administrations. As it is used to encode a wide range of resources, it can also ease the codification of interfaces, thus, helping to ensure the availability of internal information systems from public administrations.

- **Reusability**

  UTF-8 is a sector agnostic specification which means that can be used in any business domain requiring data encoding. It is also independent of any specification and can be implemented without dependencies with technologies or platforms.

- **Technological neutrality and data portability**

  UTF-8 allows for partial implementations since it's a variable-length encoding scheme that can represent any Unicode character using one to four bytes. Systems that use UTF-8 encoding can choose to implement only a subset of the Unicode character set, but this may limit their ability to handle text data from different languages and scripts. It also can be extended for multiple purposes, and be used for custom encodings, although it does not allow customisation. It is also independent of any specification and can be implemented without dependencies with technologies or platforms.

***The specification partially supports the principles related to generic user needs and expectations*:**

- **User-centricity**

  The data and information encoding is a key point in data exchange and use. In addition, the adoption of the specification as an encoding format fosters the data portability and reuse of information between administrations thus, supporting the implementation and evolution of European public services.

- **Inclusion and accessibility**

  The purpose of UTF-8 is not related to e-accessibility. Therefore, this criterion is not applicable to this specification.

- **Privacy**

  UTF-8 is neither related nor is a component of any data privacy mechanism.

- **Security**
  UTF-8 itself does not guarantee the authenticity and authentication of the agents involved in data transactions. Authenticity and authentication can be addressed through separate mechanisms, such as digital signatures, encryption, and authentication protocols. Same rationale applies to the protection of information and the provision of access control mechanisms, as additional security measures can always be implemented in top of the specification to guarantee security features. Moreover, UTF-8 encoding can enable data processing accuracy by ensuring that characters from a wide range of scripts are correctly represented and processed.

- **Multilingualism**
  UTF-8 supports many languages and can accommodate pages and forms in different cases including a mixture of these languages. The fact of allowing encoding with several languages allows the data consumption by the different linguistic groups that form the EU. Therefore, UTF-8 fosters the delivery of European multilingual services.

The specification supports the foundation principles for cooperation among public administrations:

- **Administrative Simplification**
  The use of UTF-8 encoding can help to simplify the delivery of public services as well as enable digital service delivery channels by supporting multiple languages and scripts, streamlining and simplifying data processing and exchange, thus, facilitating integration with international standards and systems and improving accessibility for users who speak different languages.

- **Preservation of information**
  The purpose of UTF-8 is not related to long-term preservation of electronic records. Therefore this criterion is considered not applicable to this specification.

- **Assessment of effectiveness and efficiency**
  There have been found some studies assessing UTF-8 in terms of effectiveness. For example, "A case study in SIMD text processing with parallel bit streams: UTF-8 to UTF-16 transcoding[6]" or a study on the influence of UTF coded data in HDB-3 operation efficiency[7].

2.2. EIF Interoperability Layers

The interoperability model which is applicable to all digital public services includes:

- Four layers of interoperability: legal, organisational, semantic and technical;

[6] A case study in SIMD text processing with parallel bit streams: UTF-8 to UTF-16 transcoding:

https://dl.acm.org/doi/abs/10.1145/1345206.1345222

[7] Compatibility of UTF-8 Encoding System to HDB-3 Scrambling Method:

https://koreascience.kr/article/JAKO201322045164159.page

- A cross-cutting component of the four layers, ‘integrated public service governance’;
- A background layer, ‘interoperability governance’.

The Specification partially supports the implementation of digital public services complying with the EIF interoperability model:

- **Interoperability governance**
  UTF-8 is associated with EIRA[8] ABBs in the EIRA Library of Specifications (ELIS)[9]. More specifically, UTF-8 is already associated with the Controlled Vocabulary, Data, Data Model, Data Syntax, Forms Structure, Hash Code and Metadata ABBs from the EIRA Library of Interoperability Specifications (ELIS) Semantic view. The specification is recommended by 10 Member States and is currently being used in European cross-border projects such as ESCO[10]. Moreover, there can be found tools for the validation for its implementation in Github[11].

- **Legal Interoperability**
  UTF-8 is not a European Standard in the sense of a formal standard developed and recognized by the European standards organizations.

- **Organisational interoperability**
  Although the purpose of UTF-8 is not related to the modelling of business processes, it can facilitate organizational interoperability agreements by enabling the exchange and processing of data in different languages or scripts. It can also help to promote the adoption of international standards or frameworks for interoperability, which can improve the compatibility and interoperability of different systems or organizations.

- **Semantic Interoperability**
  The Joinup platform holds several discussion forums[12] about the implementation and usage of UTF-8.

[8] EIRA: https://joinup.ec.europa.eu/collection/european-interoperability-reference-architecture-eira/solution/eira/release/v500

[9] ELIS: https://joinup.ec.europa.eu/collection/common-assessment-method-standards-and-specifications-camss/solution/elis/elis-dashboard

[10] ESCO vocabulary: https://ec.europa.eu/esco/api/doc/esco_api_doc.html

[11] UTF-8 validator in Github: https://github.com/digital-preservation/utf8-validator

[12] UTF-8 discussion forums in Joinup: https://joinup.ec.europa.eu/search?keys=utf-8&sort_by=relevance&f%5B0%5D=type%3Adiscussion

3. Assessment Results

This section presents an overview of the results of the CAMSS assessments for **UTF-8** The CAMSS "Strength" indicator measures the reliability of the assessment by calculating the number of answered (applicable) criteria. On the other hand, the number of favourable answers and the number of unfavourable ones is used to calculate the "Automated Score" per category and an "Overall Score".

Category	Automated Score	Assessment Strength	Compliance Level
Principle setting the context for EU actions on interoperability	100/100 (20%)	100%	Seamless
Core interoperability principles	1620/1700 (92%)	100%	Seamless
Principles related to generic user needs and expectations	980/1200 (88%)	58%	Seamless
Foundation principles for cooperation among public administrations	500/500 (88%)	80%	Seamless
Interoperability layers*	920/1000 (67%)	100%	Seamless
Overall Score	3520/3900 (90%)[13]	87%	

**The technical interoperability layer is covered by the criteria corresponding to the core interoperability principle ''Openness''.*

With an 87% of assessment strength, this assessment can be considered representative of the specification compliance with the EIF principles and recommendations.

The Overall Automated Score of 90% (3520/3900) demonstrates that the specification supports the European Interoperability Framework in the domains where it applies.

[13] See the "results interpretation" section of the CAMSS Assessment EIF Scenario Quick User Guide:

https://joinup.ec.europa.eu/collection/common-assessment-method-standards-and-specifications-camss/solution/camss-assessment-eif-scenario/results-visualisation-and-interpretation