United Nations
Educational, Scientific and
Cultural Organization

ICANN

Infoterm

# Internet Governance Glossary (IGG) Methodology

Prepared by International Information Centre for Terminology (Infoterm)
In collaboration with:
United Nations Educational, Scientific and Cultural Organization (UNESCO)
Internet Corporation for Assigned Names and Numbers (ICANN)

Photo front page: © seewhatmitchsee

The designations employed and the presentation of material throughout this publication do not imply the expression of any opinion whatsoever on the part of UNESCO concerning the legal status of any country, territory, city or area or its authorities, or concerning the delimitation of its frontiers or boundaries.

Typeset: UNESCO
Printed by: UNESCO

# I. ACKNOWLEDGEMENTS

Editorial and advisory Committee:

Baher Esmat (ICANN)
Christian Galinski (Infoterm)
Blanca Stella Giraldo (Infoterm)
Irmgarda Kasinskaite-Buddeberg (UNESCO)

Financial Support for this report was provided by UNESCO's Communication and Information Sector, Knowledge Societies Division and the Internet Corporation for Assigned Names and Numbers (ICANN) and implemented in close collaboration with the International Information Centre for Terminology (Infoterm).

We are particularly grateful to Mr Christian Galinski, Director and Ms Blanca Stella Giraldo, Deputy Director of the International Information Centre for Terminology (Infoterm), who developed the methodology for the preparation of the IG glossary and provided recommendations to the group of experts gathered at UNESCO's Headquarters from 22 to 24 April 2015 to validate the "Glossary of the Internet Governance Terms in Arabic" and to Mr Baher Esmat, Vice President, Middle East at ICANN, Egypt for his support, guidance and cooperation.

The document benefitted from significant and constructive comments received from the following experts during the international expert group meeting organized by UNESCO in April 2015:

Mr Aqel M. Aqel, Senior Consultant, Ministry of Higher Education, Riyadh, The Kingdom of Saudi Arabia; CISA Coordinator and Research Director, ISACA

Ms Nadira AlAraj, Board Member, Internet Society, Chapter in Palestine (ISOC)

Mr Khalid Al- Bulushi, Project Specialist, Digital Society Development Division, Information Technology Authority, Oman

Mr Walid Al-Saqaf, Postdoctoral Researcher, Media Studies Department, Stockholm University, Sweden, Yemen

Ms Hanane Boujemi, Programme Coordinator Internet Governance MENA region Hivos West Asia, The Netherlands, Morocco

Mr Mohamed El-Bashir, Manager of Internet, Numbering and Interconnection, ictQATAR, Qatar

Mr Ahmad Gharbeia, ICT Team Leader and Consultant, Arab Digital Expression Foundation, Egypt

Mr Abdelaziz Hilali, Chair of AFRALO, African Regional At-Large Organization/ICANN;
Chair of the Mediterranean Federation of Associations of Internet; Vice President of ISOC Morocco, Morocco

Ms Nibal Idlebi, Chief, Innovation Section, United Nations Economic and Social Commission for Western Asia (ESCWA)

Ms Manal Ismail, Executive Director, International Technical Coordination, National Telecom Regulatory Authority; Egypt Representative at GAC of ICANN

Mr Koutheair Khribi, Program Manager, ALECSO ICT Department,
The Arab League Educational, Cultural and Scientific Organization (ALECSO), Tunis

Mr Mohamad Najem, Advocacy and Policy Director, Social Media Exchange (SMEX), Lebanon

Mr Fayiz Suyyagh, Director, Tarjuman Foundation, Amman, Jordan

Mr Izzeldin Osman, Professor, Chairman of Research and Faculty Development Committee, Sudan University of Science and Technology, Programme Coordinatorregion, Hivos West Asia, UNESCO Commission National in Sudan.

The Internet Governance Glossary in English was translated by UNESCO's Arabic Language Unit, Division of Conferences, Languages and Documents and reviewed by United Nations Economic and Social Commission for Western Asia (ESCWA).

# II. ABSTRACT

Empirical evidence has shown that people whose mother tongues have not benefited from coordinated language policies and supportive tools, particularly with regard to terminology, tend to be increasingly disadvantaged in today's information and knowledge societies. Speakers of a language which lags behind in its terminology for a given domain or across many domains, risk losing the ability to communicate in specific thematic domains, lacking scientific and technical terminologies. In this context, 'language' increasingly is recognized as an issue, since the utilization of language today is highly supported by ICTs and the Internet. As a consequence, the lack of relevant terminology indirectly affects digital and knowledge divides and often forces communities to use some other, more developed foreign language for thematic domain communication. In this connection, multilingual glossaries can provide instrumental tools that greatly facilitate common understanding, communication and cooperation among various actors.

In this regard, UNESCO, ICANN and Infoterm introduce the Internet Governance Glossary Methodology; a framework to establish an Arabic glossary of Internet governance terms to support engagement in multi-stakeholder Internet governance and policy making processes by Arabic-speaking communities on various platforms, including the World Summit on Information Society (WSIS) and Internet Governance Forum (IGF).

The purpose of this document is to describe how the terms and names related to Internet governance (IG) and their descriptions have been selected and processed for the Internet Governance Glossary (IGG). Thus, it refers to the methodology of the glossary development, not to the general rational for the necessity to develop the IGG. This methodology was supposed to be state-of-the-art and – wherever possible – based on formal or non-formal standards as well as on pertinent literature. Usually – especially in prescriptive terminology work, such as terminology standardization – comparatively stable, i.e. widely agreed upon terminologies are subject to unification or standardization. In this sense, the IGG cannot be taken as prescriptive. The fact that IG terminology is partly characterized by being politically 'sensitive' – namely being perceived quite differently by major stakeholders – posed a challenge.

## SELECTION

The terms in the IGG were selected under the requirement that

- terms should belong to the most important ones for Internet Governance (IG) at a general level;
- terms should not be too technical in the respective IG dimension (i.e. sub-field of IG);
- descriptions of the terms should be user-friendly – i.e. somehow understandable by non-experts of the technical issues covered.

The proper names of presently existing major organizations, forums, networks, groups, conferences, regulations and legal instruments pertinent to IG were selected under the perspective of international or at least regional relevance on the basis of existing documentation and recommendations by experts. They are contained in Section VII of the IGG.

There are other names occurring in the IGG, which could also be considered as terms standing for individual concepts, such as the “Internet Protocol version 4” (IPv4). Such names are treated as terms.

# TABLE OF CONTENT



## III. BACKGROUND

The 'digital divide' is a term that refers to the gap between demographics and regions that have access to modern information and communication technologies (ICTs), and those that do not or only have restricted access. ICTs include the telephone, television, personal computers and the Internet, as well as all the related software and applications. Well before the late 20th century, 'digital divide' referred chiefly to the division between those with and without telephone access; after the late 1990s, the term began to be used mainly to describe the split between those with and without Internet access.

The digital divide typically exists between those in cities and those in rural areas; between the educated and the uneducated; between socioeconomic groups; and, globally, between the more and less industrially developed nations. Even within populations with some access to technology, the digital divide can be evident in the form of lower-performance computers, lower-speed wireless connections, lower-priced connections such as dial-up, and limited access to subscription-based content. A June 2013 U.S. White House broadband report, for example, showed that only 71% of American homes have adopted broadband, a figure lower than in other countries with comparable gross domestic product. Thus, the digital divide is very much a reality and particularly refers to broadband access to the Internet today.

The reality of a separate-access marketplace is problematic because of the rise of services such as video on demand, video conferencing and virtual classrooms, which require access to high-speed, high-quality connections that those on the less-served side of the digital divide cannot access and/or afford. And while adoption of smartphones is growing, even among lower-income and minority groups, the rising costs of data plans and the difficulty of performing tasks and transactions on smartphones continue to inhibit the closing of the gap.

The United Nations Educational, Scientific and Cultural Organization (UNESCO) encourages States to use ICTs to promote greater participation by citizens in democratic life. This can be achieved by:

- Using the Internet and other ICTs as tools for dialogue between citizens and the authorities;
- Integrating new and "traditional" technologies, including library services and community media; the production, adaptation, translation and sharing of local contents; and the setting up of pilot projects corresponding to different cultural contexts;
- Giving high priority to the needs of those disadvantaged and marginalized groups that are presently excluded, so that information societies be open and inclusive;
- Improving access to the benefits of the knowledge societies for women and youth;
- Extending material assistance to countries at present unable to offer access to ICTs to large numbers of their citizens.

Proponents for closing the digital divide include those who argue it would improve literacy, democracy, social mobility, economic equality and economic growth.

Intergovernmental organizations such as UNESCO and global organizations such as the Internet Corporation for Assigned Names and Numbers (ICAN), in addition to regional organizations, non-governmental organizations (NGOs) and others are highlighting the indispensable role of languages in building inclusive Knowledge Societies. Each language offers a unique testimony of its civilization's cultural genius and contributes to the world's heritage – not to mention its crucial role in building intercultural dialogue, reconciliation and peace.

Empirical evidence has shown that people whose mother tongues have not benefited from coordinated language policies and supportive tools, particularly with regard to terminology, tend to be increasingly disadvantaged in today's information society. In many cases, when a language is more or less confined to the family sphere, it starts losing its importance within the professional and international community. In other words, speakers of a language which lags behind in its terminology for a given domain or across many domains, risk losing the ability to communicate in different thematic domains in their language over time. This implies that a language community whose language has not developed scientific and technical terminologies is unavoidably forced to use some other, more developed foreign language for thematic domain communication. In this connection, multilingual glossaries can provide instrumental tools that greatly facilitate common understanding, communication and cooperation among various actors.

In its early development, the Internet would often be assessed from a technological perspective. However, during the last decade, other issues, often qualified as "soft", have emerged focusing on topics such as human rights, democracy, privacy, social equity, inclusion, local content creation, interdependence, and other cultural, educational, economic and political aspects of Internet use. Such discussions have been ongoing since the creation of the Internet Governance Forum (IGF), which resulted from the two World Summits on the Information Society (WSIS in Geneva, 2003, and Tunis, 2005) and WSIS+10 Review Process and outcomes. In this context, 'language' increasingly is recognized as an issue, as the utilization of language today is highly supported by ICTs and the Internet. As a consequence, the lack of relevant terminology indirectly affects digital and knowledge divides.

Today, the international community has several multi-stakeholder mechanisms for the dialogue and implementation of solutions to Internet governance issues. The WSIS events and IGF belong to some of the major ones. To participate in international multi-stakeholder processes, countries and their national representatives need also to be equipped with language tools that facilitate understanding, cooperation and coordination. One of the latter are the rapidly growing multi-stakeholder partnership mechanisms related to the governance of the Internet. Under this aspect, Arabic-speaking countries and communities currently have limited opportunities to be fully engaged in constructive dialogue and joint action in multi-stakeholder processes due to the inadequate use of supportive language tools. Hence, solutions are needed to strengthen technical terminology in the Arabic language in order to facilitate the dialogue on the use of Arabic on the Internet, in an effective, efficient, and coordinated manner.

Partners involved in the IGG project commonly agree that the Internet should serve all people around the world. The technological advances of the Internet open up vast opportunities to access, preserve, create, and share information and knowledge. When information is shared on the Internet, it immediately becomes available to a large audience and can have a global impact. However, it is important to ensure that information and knowledge are made equally available not only to the world's most dominant languages, but also to lesser-used languages. The IGG project aims to provide Arabic-speaking countries and communities with a glossary on Internet governance (IG) along with a formal generic methodology for the creation or harmonization of terminology based on international standards.

The partners involved in the preparation of the project proposal were keen on methodological assistance in developing a specific language tool entitled 'A Glossary of IG terms' in English and to localize it into Arabic by working in close collaboration with Arabic-speaking and international experts and organizations involved in IG, IT and linguistics. The aim is to facilitate involvement and participation of Arabic-speaking countries and communities in the multi-stakeholder processes, particularly on Internet governance issues. The tool would be used by Arabic-speaking countries and communities around the world, particularly national professional organizations working in the field of language policy development, stakeholders involved in Internet governance (IG) and other multi-stakeholder processes, and Arabic speakers in general. It can also be seen as an initiative to comply with the need for continuous terminology planning, institutional capacity building and effective coordination mechanisms at regional and country levels.

At the end of the IGG project, it is expected that Arabic speaking countries will use the glossary of Internet governance terms developed in a collaborative, accurate and multi-stakeholder way;

- To help formulate joint agendas, contribute to decision making processes, communicate more effectively and encourage their engagement in multi-stakeholder IG mechanisms;
- To participate in all kinds of multi-stakeholder processes;
- To build institutional capacities of and coordination mechanisms among regional and national organizations working on language issues with a view to contributing to the promotion of and access to the Arabic language on the Internet.

It is anticipated that once a coherent, harmonized glossary of Internet governance terms is in place, new terms and their descriptions could be adopted and used by policy makers, public administrations, media, educational institutions, and the public at national, regional and international levels. To ensure the widest dissemination of the glossary, inclusive consultation and feedback mechanisms such as online and face-to-face discussions and consultations are integrated into the project. Below, a list of major steps identified and followed for the achievement of the expected result is given:

1. Mapping key documents on the IG issues for the development of a draft glossary:
For the preparation of the draft glossary of IG terms in English, a significant number of major documents (more than 160) are identified and analysed as well as results incorporated in the initial list of terms on Internet governance issues.

2. Preparation of a draft glossary of the IG terms in English:
A number of experts from the field of terminology, IG and other relevant fields are consulted and contribute to the development of that initial list of terms. The draft of the IG glossary in English includes all identified IG terms.

3. Public consultations:
The draft glossary is used for public online consultations with experts representing national, regional and international professional organizations. Based on the comments and recommendations received, a pre-final version of IG terms in English is prepared, then translated and used for the adaptation and localization of the IG terms in Arabic.

4. Localization of the glossary of IG terms into Arabic and validation process: Another round of public consultations is initiated involving relevant experts and organizations. The draft glossary of the IG terms in Arabic is revised and used for the face-to-face validation meeting in Paris, France at UNESCO Headquarters.

5. Distribution:
The final e-document is published, launched and distributed by the partners during international events, such as the Internet Governance Forum, WSIS and other events.

The Internet has become way more important than just for research purposes or for the economy – it has become one of the key pillars of modern society linked to fundamental human rights (including access to information, freedom of expression), health, and education. Unlike other technologies, the Internet has 'users' rather than 'consumers'. That is why entirely profit-led models (even if clearly leading to more innovation and investment) may increase the divide between the information-rich who would be using unlimited online services with full quality, and the information-poor who might have to content themselves with useless best effort services. Nobody denies that the Internet – representing a new 'technological revolution' – has brought about and is still continuing to provide tremendous benefits to society at large. But there are also real risks.

Thus, Internet governance is a complex multi-facetted and multi-dimensional topic which touches upon technical (referring first of all to infrastructure and standardization), economic, legal, development and socio-cultural aspects, as well as those of stakeholders, a broad range of activities and the results thereof. This dynamic development of the Internet and complex nature of the topic had an impact on the methodology insofar as for instance:

- A random selection of terms and proper names did not appear to be very useful (especially when looking at existing IG related glossaries);
- Statistics of occurrences such as by means of computer-linguistic methods and tools did not prove to be suitable either.
- Newer sources often provided information not existing or described differently from those of earlier years.

This situation necessitated a highly systematic and meticulous approach from the outset.

The methodology of preparing the "Glossary of IG Terms" (IGG) while following international standards had to be adapted in order to cope with the above-mentioned situation.

1. A preselection of official documents, which have been analyzed and supplemented in close consultation with the above-mentioned cooperation partners, was taken as the basic 'text corpus' for the identification and description of the main IG terms. Based on the analysis of the nature of these documents, those promising to provide good contexts were given priority in the beginning. At a later stage, the drafted IGG entries were once more checked against the comprehensive text corpus of all documents.

2. An alphabetical list of specialized terms and proper names with their abbreviations – as mostly provided in term collections on the Internet – was not considered as promising for yielding good entries which should be mutually consistent and coherent. In addition, IGG versions in other languages – where the entries would certainly not follow the alphabetic order of the terms in English – had to be considered from the beginning. Therefore, the systematic approach chosen started off from a broad sub-division of IG dimensions as it emerged from the analysis of the preselected documents.

3. Although the IGG is supposed to follow international standards and best practices, 'definitions' in the strict sense were not considered as appropriate in view of the anticipated end users of the IGG, namely well informed laypersons, rather than highly specialized experts. The definitions are supposed to be 'user-friendly', not scientifically stringent. Therefore, a form of condensed 'description' of the concepts represented by the terms of the IGG was chosen. Special care has been taken to formulate the descriptions in such a way that the conceptual interrelation between the IGG entries became clear.

The analysis of the contexts of the IGG term candidates in the preselected documents revealed that most occurrences did not provide sufficient explanations or were not coherent or even contradictory. Therefore, additional sources – including research papers, online computer encyclopedia, Wikipedia, websites of organizations, etc. – have been consulted. As a result of this process, nearly each IGG entry constitutes an 'adaptation/amalgamation' of a number of occurrences from different sources.

# IV. GENERAL INTRODUCTION

The purpose of this document is to describe how the terms and names related to Internet governance (IG) and their descriptions have been selected and processed for the Internet Governance Glossary (IGG). Thus, it refers to the methodology of the glossary development, not to the general rational for the necessity to develop the IGG.

The terms were selected under the requirement that

- terms should belong to the most important ones for Internet Governance (IG) at a general level;
- terms should not be too technical in the respective IG dimension (i.e. sub-field of IG);
- descriptions of the terms should be user-friendly – i.e. somehow understandable by non-experts of the technical issues covered.

The proper names of presently existing major organizations, forums, networks, groups, conferences, regulations and legal instruments pertinent to IG were selected under the perspective of international or at least regional relevance on the basis of existing documentation and recommendations by experts. They are contained in Section VII of the IGG.

There are other names occurring in the IGG, which could also be considered as terms standing for individual concepts, such as the "Internet Protocol version 4" (IPv4). Such names are treated as terms.

The methodology applied can be considered as state-of-the-art and – wherever possible – based on formal or non-formal standards as well as on pertinent literature. Usually – especially in prescriptive terminology work, such as terminology standardization – comparatively stable, i.e. widely agreed upon terminologies are subject to unification or standardization. In this sense, the IGG cannot be taken as prescriptive. The fact that IG terminology is partly characterized by being politically 'sensitive' – namely being perceived quite differently by major stakeholders – posed a challenge.

# 1 PREPARATORY CONSIDERATIONS

## 1.1 DIFFERENT KINDS OF TERMINOLOGY WORK FROM THE POINT OF VIEW OF METHODOLOGY

The methodology to be chosen for terminology work depends first of all on the following questions:

1. Is the terminology of the domain or subject highly stable or undergoing rapid change?

→ The IG terminology is in parts a fast emerging rather than stable terminology.

2. Who belongs to the target audience of the terminology?

→ The IGG is geared towards non-experts of some or most dimensions of content covered by the IGG, which means:

- Terms to be selected should not be too technical in each of the dimensions covered.
- Descriptions should be understandable as much as possible by non-experts of the respective dimension.

3. What is the purpose of the IGG?

→ The IGG is aiming at establishing transparency of the meaning of the terms used in IG within languages and between languages, and therefore:

- IGG entries should be cross-referencing each other in order to establish a semantic 'context' for a major part of IG – especially for the higher conceptual levels.
- Descriptions for the terms should be formulated in such a way that they are not too detailed and not too technical, and yet creating the above-mentioned semantic 'context' for better understanding.
- While preparing the IGG entries in English, due consideration should be given from the very beginning to the need of translating – or rather 'transcreating' – them later into other languages.

From the above it became clear that:

→ a purely 'descriptive' approach would not work, because it would create huge amounts of information when comparing the meanings of terms in various contexts of many documents which would confuse the target audience,

↳ *Nevertheless – as in any kind of terminology work – it needs a descriptive phase at the beginning of a terminology project.*

→ a purely 'prescriptive' approach – which needs to be based on some 'authority' for prescribing the wordings – would not work either, because it would necessitate a highly systematic approach harmonising each concept represented by one or more terms,

↳ *Nevertheless – given the constraints from the outset – a certain 'prescriptive' element was unavoidable (not least with respect to data administration and layout).*

In any case, even within 'descriptive' approaches there is a range of degrees of descriptiveness, and within 'prescriptive' approaches one can find a range of degrees of prescriptiveness.

For pragmatic reasons, certain decisions concerning the workflow and the presentation of the data have to be taken in any of the approaches mentioned. As there are several similar or complementary state-of-the-art models for all kinds of terminology work – including for glossaries containing specialized terminology – the methodology, workflow and presentation of data for the IGG follow the basics of these models, though slightly adapted for the IGG purpose. Thus, a mixed approach – combining descriptive and prescriptive elements – has been chosen. This approach is also in line with specialized lexicography approaches which focus on the terminology occurring in the specialized communication of a domain.

## 1.2 DIFFERENT KINDS OF WORKFLOW DESIGN AND PLANNING

Workflow design may differ depending on the methodological approach and given constraints. However, usually the design of the workflow observes the following major phases (according to ISO 15188:2001):

- Preparation phase: considering feasibility, framework and specifications;
- Design phase: considering project leadership and project planning with all its aspects (incl. the use of tools);
- Implementation phase;
- Review, evaluation and verification phase;
- Final phase (evaluating the results and concluding the project).

In real terminology work practice, depending on the purpose and the resources (in terms of human resources and technology) available, each project requires a more or less adapted workflow.

Therefore, the partners –according to best practice– duly adapted the rules of major International Standards of ISO/TC 37 "Terminology and other language and content resources", such as

- ISO 10241-1:2011 Terminological entries in standards — Part 1: General requirements and examples of presentation
- ISO 10241-2:2012 Terminological entries in standards — Part 2: Adoption of standardized terminological entries
- ISO 15188:2001 Project management guidelines for terminology standardization

Duly taking into account some rules and information from:

- ISO 12620:2009 Terminology and other language and content resources — Specification of data categories and management of a Data Category Registry for language resources

- ISO 26162:2012 Systems to manage terminology, knowledge and content — Design, implementation and maintenance of terminology management systems.

As the following figures are showing, intermediary results may have to be referred back to a previous phase, if necessary.

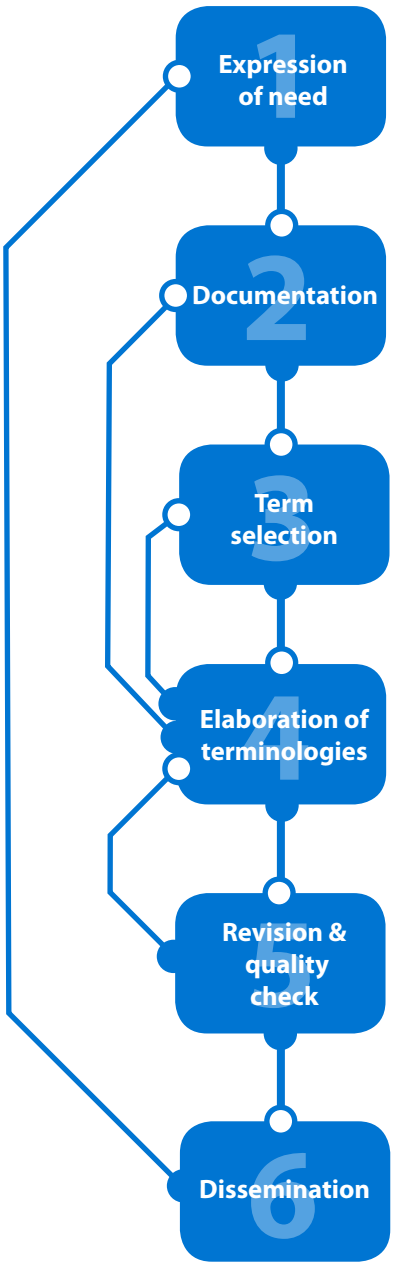


*Fig. 1: Main activities of terminology work*

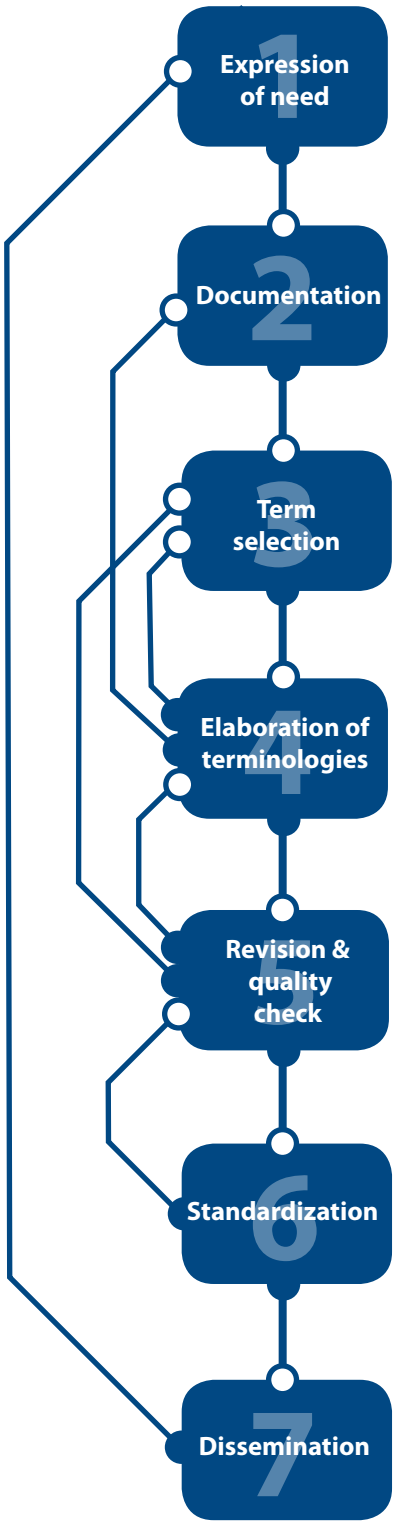


*Figure 2: Complete terminology workflow with standardization*

Being aware of the fact that the IGG would comprise also entries representing proper names of organizations, forums, networks, groups, conferences, regulations and legal instruments pertinent to IG, as well as extended terminological entries falling into the category of "terminological phraseology", the approach outlined in the article "Methodology for the definition of a glossary in a collaborative research project and its application to a European Network of Excellence" (Velardi et al.) was taken as a reference for the state-of-the-art of glossary making from the pragmatic point of view of specialized lexicography using ontology tools:

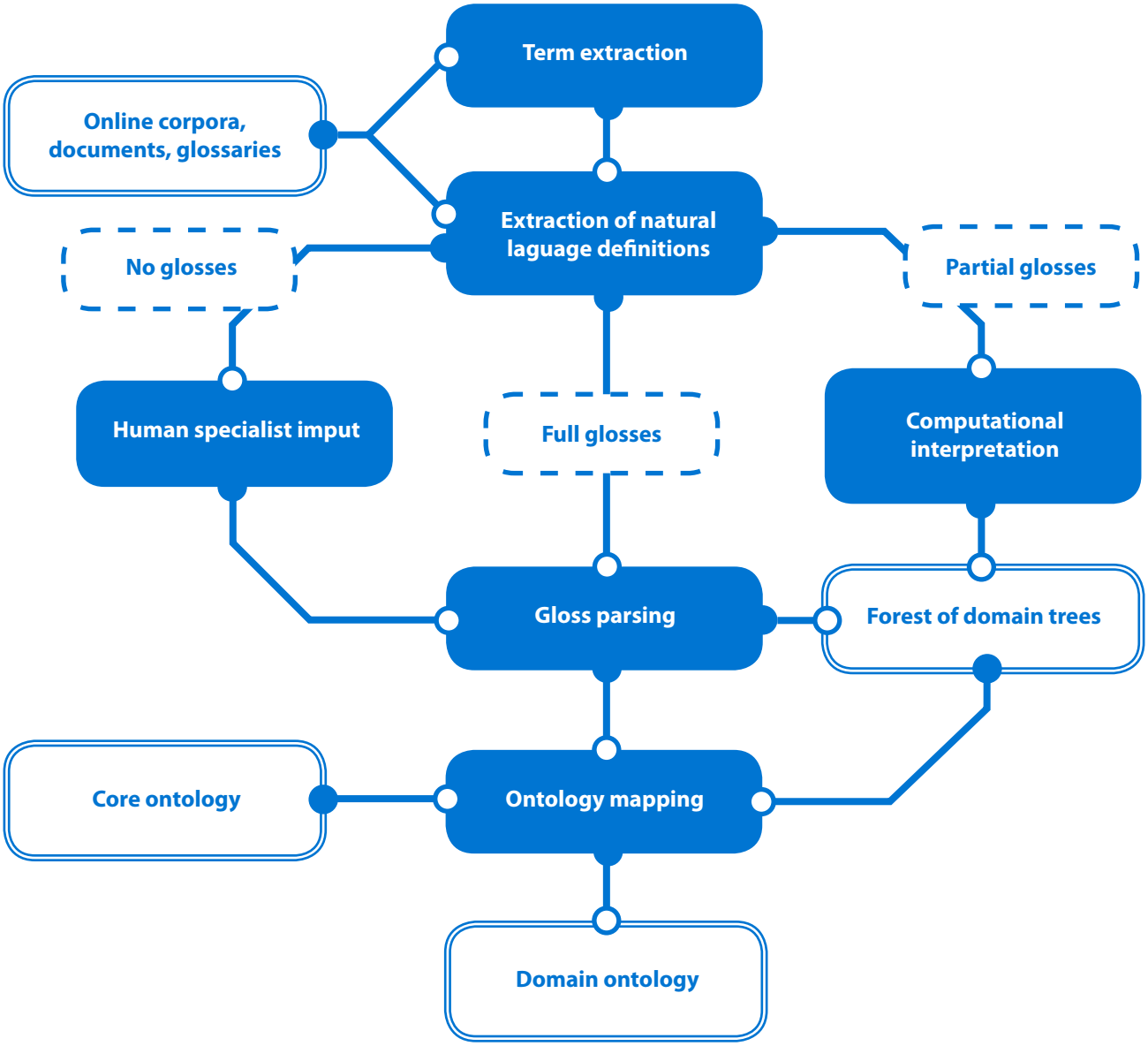


*Figure 3: The workflow model of ontology-based glossary development (simplified)*

Given the nature of the IGG project, existing best practices were adapted and combined while taking advantage of the tools that suited the characteristics of the documents available. The "Methodology for the development and validation of the glossary on the Internet Governance" (IGG methodology) was drafted on the basis of the following main stages:

**1st stage:** Definition of the purpose of the glossary (IGG workflow step 1)

**2nd stage**: Selection of the documentation provided and preparation of a first list of IG terms with contexts (IGG workflow steps 2~6)

**3rd stage:** Pre-evaluation phase (IGG workflow steps 7~15)

**4th stage:** Establishing a draft glossary taking into account the input from the pre evaluation phase and using a term extraction tool (IGG workflow steps 16~22)

**5th stage:** Consultation of a wide audience of experts for the finalization of the glossary (IGG workflow steps 23~26)

While the 'generic' methodology chosen for the IGG is state-of-the-art and – wherever possible – based on formal or non-formal standards as well as on best practice, each implementation has to take into account each environment with its specific requirements and constraints. Thus, the IGG workflow was adapted in the course of its development.

The steps to be followed under each stage were carried out sometimes in parallel whereby avoiding repetition of certain steps as much as possible – not least with a view to time constraints. The outcome was a sequence of 26 steps under the commonly applied 5 stages. For stage, the purpose was defined envisaging the final outcome of the IGG in consultation with experts and representatives of stakeholder organizations. In addition, a first selection of documents pertinent to IG was provided for the consultations. For stages 2 to 4, a pragmatic approach was chosen subdividing them into detailed steps (as manifested in the workflow chart below – see Figure 4).

The originally conceived workflow was implemented step-by-step in the course of developing the IGG project and proved to be sufficiently flexible when modifications were necessary. It included the following steps (each of them covering a set of actions):

1. **Collection of documents pertinent to Internet Governance (IG):**
About 30 experts collected almost 150 potential sources on which to base the terminology work for the IGG. These documents also comprised lists of suggested term candidates for the glossary and already existing more or less pertinent glossaries.

2. **Consolidated selection of documents to be analysed:**
About 100 documents were chosen to constitute the reference list.

3. **3. Categorization documents according to defined criteria:**
The textual documents were separated from the glossaries and further pertinent sources were added. Then the documents were renumbered, coded and categorized according to:

   - type of document (i.e. sort of text, such as report, article etc.),
   - assumed suitability for term extraction (categorized as highly suitable, less suitable and to be neglected at least during the phase of term extraction). (for detailed information see para 1.3).

4. **Outlining the glossary methodology:**
(in parallel to step 3) The basics of the methodology to be applied were drafted taking into account the circumstances of the IGG project in terms of capacities and time framework.

5. **Selection of IGG term candidates with contexts:**
Based on the texts classified as A and B documents (see para 1.3) and the state-of-the- art literature, the term candidates with contexts and annotations were selected and decided on five IG dimensions to group the terms according to sub-themes.

6. **Document preparation for testing term extraction tools (TET):**
(in parallel to step 5) several text corpora were established under the "ProTerm" software so that any useful subset of documents can potentially be analysed for selected term candidates. (For detailed information see para 3.2)

7. **Providing a draft list of term candidates in pre-evaluation format:**
As the data collected in step 5 proved to be very voluminous, a format for starting with the pre-evaluation process was provided. This included suggestions for simplifying the descriptions of the terms.

8. **Preparation of a comment template:**
(in parallel to step 7) a template for gathering comments through electronic communication was developed.

9. **Proposal for drafting the descriptions** (in parallel to step 7)

10. **Consolidating the list of term candidates for the IGG:**
In consultation with the partners, the first draft of IGG was used for pre-evaluation.

11. **Selecting and adapting a term extraction tool (TET):**
(in parallel to step 10) The term extraction (or better, terms and names identification) tool "ProTerm" was selected and adapted for the IGG purpose in such a way that all or any useful subset of the documents can potentially be analysed for selected term candidates. Several text corpora (according to the document categories in step 3) were established under this software. (For detailed information see para 3.3)

12. **Pre-evaluation by 10 experts selected and consulted by telecommunication:**
10 experts were consulted on the selection of IGG term candidates based on the consolidated first draft of the IGG. The experts selected terms from the term candidates, de-selected others and added a few new term candidates. Some experts also commented on the dimensions and rearranged some of the term candidates according to the dimensions.

13. **Pre-evaluation results:**
The pre-evaluation results were considered in the further work on the IGG.

14. **Checking and processing pre-evaluation data:**
The data was checked and integrated into a preliminary table of contents for the future IGG. The de-selected term candidates were kept for the time being in a separate list.

15. **Analysing the integrated table of categorized term candidates:**
(in parallel to step 14) The re-compiled term candidates were analysed and integrated into a preliminary table of contents for the future IGG.

16. **Preliminary consolidation of the re-compilation results:**
A template for the comments was prepared a to facilitate the assessment of recieved comments

17. **Consolidating the IGG methodology** (in parallel to step 16)

18. **Discussion of the draft table of contents:**
The partners discussed the draft table of contents and provided late pre-evaluation comments.

19. **Finalized selection of categorized terms:**
On the basis of the discussions, the selection of term candidates for the IGG at that stage was finalized.

20. **Drafting user-friendly descriptions:**
In the course of the preparation of user-friendly descriptions, several actions were carried out that contributed to reach the objective such as:

- putting the selected terms into a thematic order taking into consideration the given recommendations and their pertinence,
- reducing the volume of the descriptions as well as the number and volume of the notes,
- combining a couple of entries according to the consolidation of the description and the context extracted from ProTerm (e.g. eSignature and digital signature, cybercrime…),
- separating a couple of entries (e.g. TCP and IP),
- adding a couple of entries (e.g. Internet),
- providing cross-references (highlighted in red and bold --> not yet complete, and not yet in the final layout),
- adding sources by using the term identification and extraction tool ProTerm,
- NOT adding explanatory text to the organizations and other names, such as conventions/treaties,
- taking-off the de-selected terms and putting them into a different file (for archiving purposes)

**21. Computer-assisted comparison of contexts:**
(in parallel to step 20) ProTerm proved to be very useful in identifying significant contexts as material for the descriptions, clearly keeping track of the source(s), etc.

**22. Editing the IGG methodology:**
(in parallel to step 20) In the course of stage 4 the IGG methodology was pre-finalized.

**23. Consolidation of the IGG entries:**
A total revision and edition of the entries was carried out integrating the input received and the pertinent context extracted to complete the descriptions. For the validation meeting, all material was prepared as well as a list of the terms in alphabetic order (i.e. as an index). A note to the layout of the IGG entries was added after the table of contents.

**24. Translation of IGG entries into Arabic:**
Being aware of the fact that many terms related to IG and ICT have been borrowed mainly from English or French into Arabic, the document was translated into Arabic.

**25. Face to face validation meeting with experts and stakeholders**

**26. Integration of last input and final consolidation of the IGG:**
A final Arabic-English version of the IGG with about 200 terms (under seven dimensions) comprising about 50 proper names of major organizations, forums, networks, groups, conferences, regulations and legal instruments, was consolidated following the input and issues or recommendations discussed.

In the course of formulating the descriptions, special care had been taken to provide cross-references to conceptually related entries. This and the harmonization of style, layout etc. of all entries necessitated a meticulous adaptation of the formulations.

It is important to recognize that a final process took place after step 26:

- to agree on all modifications to the Glossary
- to finalize the description of the methodology.

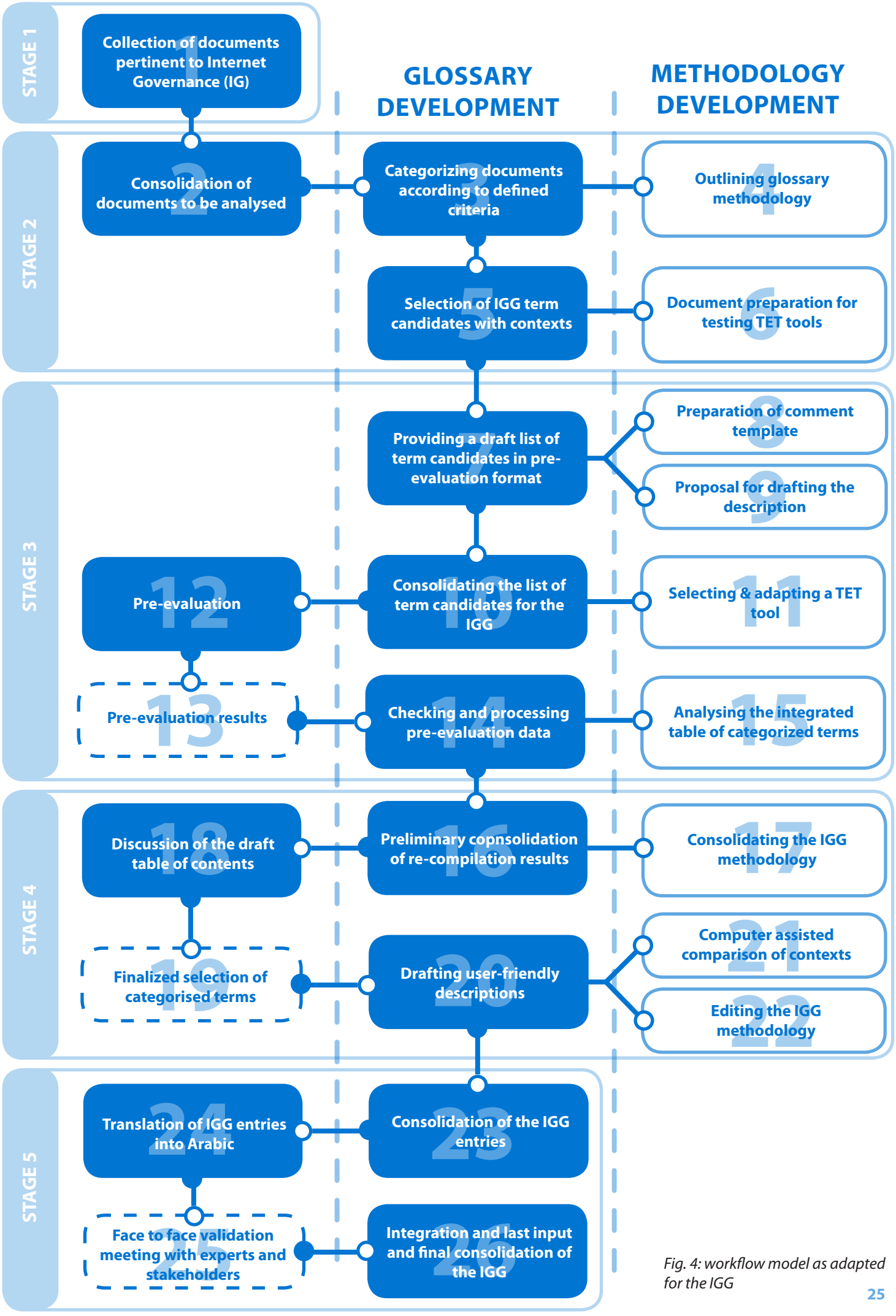


*Fig. 4: workflow model as adapted for the IGG*

## 1.3 DOCUMENTATION CONSULTED FOR THE PREPARATION OF THE IGG

In general, any glossary may include a list of terms and associated definitions or descriptions (and/or proper names) which either have been selected from already existing sources or by experts, or developed by the author(s) of the glossary. Common challenges encountered by glossary developers are:

- Many existing glossaries are too "flat", taking for granted that people/target groups understand the "context" of the individual glossary entries or the respective sources.
- Many glossaries have entries with terms and or proper names ordered alphabetically which require 'understanding' of terms out of context. However, for the aim of the IGG, the grouping of the terms by dimensions and/or fields or subfields provided the necessary context for the comprehension and translation of the glossary.
- Additional sources often contain quite contradictory information when comparing entries of related content.

Therefore, the main focus had to be laid on using the documents selected for the IGG project. In any case, nearly each IGG entry had to be constructed/adapted from several sources: including existing glossary entries, texts/documents and other sources of the domain. This is due to several issues such as: many documents do not describe the terms themselves; they only mention or refer to them in a shallow way either because the meaning is taken for granted or it is a concept sufficiently known to the target audience. Anyhow, the lack of precise context, conceptual harmonization or lack of updated information made it necessary to adapt, complete or reformulate most of the entries' description.

Defining criteria to evaluate the documents with a view to establishing text corpora with them was a first complex task. The list of documents (together with the files) was analysed and documents categorized. Additional documents have been added in the course of that task. A first selection of 107 documents and three other resources were analysed taking into account the following criteria:

- Pertinence of the information provided in the documents not only in relation to the field but also in terms of conceptual context,
- Authoritative documents that serve as a guidance for users,
- Key documents such as handbooks and guidelines,
- Existing glossaries were also considered, mainly to contrast with the selected term candidates and to consolidate term descriptions and explanations,
- Date of issue,
- Origin – especially issuing authority – of the documents, etc.

To facilitate further processing throughout the project, the documents were first labeled not only with unique numbers, but also by adding a mnemonic short form of the document title plus a code for the nature of the document, such as:

- "ar" (annual review)
- "acro" (contains acronyms)
- "bapa" (background paper)
- "bripa" (briefing paper)
- "CoEu" (Council of Europe)
- "d" (draft)
- "dipa" (discussion paper)
- "dwp" (draft working paper)
- "EuCo" (European Commission)
- "glo" (glossary)
- "HB" (handbook)
- "hist" (history related)
- "IS" (Internet Society)
- "ov" (overview)
- "pol" (policy related document)
- "rep" (report)

This combined reference codes were used throughout the compilation of the IGG and replaced at the final stage of the project by conventional bibliographical referencing in Annex 1 of the IGG.

All documents were categorized for further selection based on the pertinence to the project as follows:

**1.** **Assigning a "document qualifier" classifying the documents from the point of view of potential relevance/usefulness of the terminology contained:**

T = contains relevant/pertinent terminology
F = document is of a fundamental nature (in general or with respect to one of the topics of IG)
P = document is of a political nature (preferably at international level in general or with respect toone of the topics of IG)
I = document contains lots of ICT terminology (with respect to IG in general or to one of the topics of IG)
D = draft or report or summary or (lower level) recommendation
O = old document (with respect to the purpose of our project, which does not preclude that an older version in some cases may be more pertinent=not old than a newer document classified as old)

**2.** **"Priority" from a terminology point of view**

A = highly relevant in terms of terminology contained
B = relevant to some extent in terms of terminology contained
C = low relevance in terms of terminology contained

The above categorization by document nature, pertinence for term selection as well as potential relevance/usefulness of the terminology contained proved to be useful to find the most effective method to start the IGG project. Following the above pertinence categories:

- 11 documents were categorized as definite key documents (possibly optimal starting points for the term extraction); ultimately 14 definite key documents were identified.

- Further 22 documents were categorized as additional key documents, which still needed to be checked, whether they really contain more suitable terminological information (in terms of defining contexts) than other key documents.

The categorization of the documents was based on a systematic approach from the outset. First, each document underwent reading strategies by terminology experts, such as scanning and skimming, to categorize them. In the course of this process, documents were compared and the categorization aligned, when, for instance, they belonged to the same type. In some cases, however, the categorization was heightened or lowered, compared to similar documents, depending on their pertinence for the IGG project. In some other cases, a categorization was not possible, because the document did not contain pertinent/relevant terminology or did not add new information.

All in all, the term extraction process started with a selection of about 25-30 documents. However, this analysis and categorization did not mean that the other documents were to be disregarded, but it was geared towards making the whole endeavor efficient/manageable from the beginning. In this connection, particular care was necessary not to lose focus: i.e. not to get too much involved in related aspects, such as ICT or economic terminology. Otherwise, the IGG would easily go far beyond the core IG terminology. Nevertheless, ultimately more than 150 sources were used for extracting and evaluating terms.

While starting term selection manually, term extraction software was identified and tested. Most of the systems available had serious disadvantages. (See: part 3).

## 1.4 TERM CANDIDATES AND THEIR CONTEXTS IN THE DOCUMENTS

In the case of the IGG, the following observation could be made concerning the kind of information in the context of term candidates:

- In most occurrences, term candidates appeared without defining context in general.
- Even in the documents assumed to be more suitable for term extraction, term candidates mostly appeared with insufficient context, which necessitated:
  - either to combine the contexts of several occurrences,
  - or to consult additional sources.
- Many terms were not consistently used over the broad variety of texts.

In the beginning, contexts were collected as much as thought pertinent, and an attempt for first draft definitions/descriptions was made. In a later phase, the descriptions had to be simplified in view of the target audience of the IGG while checking again the most significant occurrences with the help of a term/name extraction tool. In the formulation of the descriptions of the terms in the IGG entries, a preference was given to general understandability and user-friendliness. It is important to highlight that statistical information (e.g. about word frequency) proved not to be the most useful criteria to select a term or to extract proper descriptions, sometimes high frequency rates were due to layout features or references but not to meaningful information.

# 2 METHODOLOGY OF DRAFTING INDIVIDUAL IGG ENTRIES

There are basically two kinds of IGG entries:

- Terms (or names such as IPv4) each representing a more general or more specific, however, closely IG related concept.
- Proper names of organizations, forums, networks, groups, conferences, regulations and legal instruments pertinent to IG.

Because most of the terminological entries listed in the draft IGG were selected from different sources related to the IG subject area, decisions for a unified, systematic approach for handling terms and descriptions has been necessary from the outset (which was positively confirmed in the course of work on the IGG). However, this unified approach does not mean that each individual entry is uniformly presented in the same way. The 'usage' status of each term or name for instance has been indicated on a case by case basis, since not every term has synonyms or abbreviated forms. Many entries have two or more synonyms; some are full forms accompanied by abbreviated forms etc.

The IG glossary includes a number of terms or names which are relatively new, nevertheless widely used without common agreement/standardized definition within the IG communities. Other terms are widely used and well defined, and therefore, can be considered as 'standard'. Some infrequent terms may be important for clarification, although they are very rarely used. Term extraction tools usually cannot differentiate between important/significant occurrences and non-significant ones. Therefore, statistical evidence only was not enough to determine the importance of a term or to identify a good description or explanation from the context.

The standards-based systematic – while adapted – approach refers first of all to:

- the method for compiling the IGG, including the methodology of managing the great amount of documents;
- the macrostructure and microstructure of the IGG;
- the layout of the entries, including the use of symbols.

It comprises – whenever possible unified – rules for the structure of the IGG, term usage and semantic relations, orthographic and other writing conventions, descriptions of the entries.

In line with time-honoured practice in specialized lexicography and terminology work, it was decided to present each IGG entry (each comprising a.) a term together with synonyms or abbreviated forms, as well as related terms, if existing, and b.) a description followed by a note, if necessary) as a separate entity. The entries are grouped under the respective IG dimension which is also providing a certain conceptual 'context'. This 'mixed order' (i.e. non alphabetic order) became prominent in terminology standardization – first in the form of the Electrotechnical Vocabulary (IEV) at the beginning of the 20th century and later in the majority of terminology or vocabulary standards. It also became best practice

in some highly technical specialized lexicography endeavours, such as in the series of the multilingual Illustrated Technical Dictionaries compiled by Alfred Schlomann in the 1990's.

A semi-systematic approach (resulting in a mixed order of entries) to the ordering of glossary entries in a larger glossary is proving times and again its suitability for several reasons:

- Broadly dividing the glossary into sub-themes helps to select and prioritize a proper number of terms.
- Grouping the entries according to their conceptual relatedness helps to establish cross-references, fine-tune descriptions, avoid inconsistencies etc.

Therefore, the mixed (or semi-systematic) order was chosen for arranging the entries in the IGG.

## 2.1 MICROSTRUCTURE OF THE IGG ENTRIES

The microstructure of the IGG entries was considerably simplified which is justified in view of the target audience.

1) Term or name level (with examples):

| Field | Explanation |
|---|---|
| 2.13 | Number of the entry in the IGG |
| **cloud computing** | Term (preferred terms coming first) or name **is written in bold letters**. Each term/name or its abbreviated form is entered in a new line. |
| **Internet operator** | Terms are written with lower case letter at the beginning, unless they are names beginning with a capital letter or with a letter or character subject to other conventions. |
| RT: TCP/IP<br><br>NOT: carrier | A synonym: may have different usage status, such as preferred (fully equivalent to the first term of the entry and, therefore, also in bold letters), admitted (in regular letters) or deprecated;<br>→ If it is a more or less related term (i.e. a quasi-synonym or a term/name commonly used), it is preceded by "RT:" (i.e. 'related term' covering also quasi-synonyms)<br>→ If it is a term/name that should not be used (e.g. a deprecated term), it is preceded by "NOT:" |

2) Description level (with examples):

| Field | Explanation |
|---|---|
| <Internet infrastructure> | <expression in angular brackets> indicates a specific domain or subject of term use to make clear that meanings of the term in other domains or subjects are excluded. |
| term largely referring to telecommunications service provider (TSP) and **Internet service providers** (ISP) | Description – sometimes definition – of the concept represented by the term or name. Cross-referenced IGG terms are **written in bold letters.** |
| NOTE: In order to provide quality service, Internet operators are using various traffic management techniques to prioritize certain traffic. | If a note is useful for understanding the concept of the entry, it is indicated by "NOTE:" There may be more than one note in a different language version of the entry. |

| | |
|---|---|
| [Wikipedia]<br><br>[Wikipedia adapted] | Indication of the source (in short form) in square brackets; the full version is listed in the "List of references".<br>Often the text of the source had to be adapted, which is indicated by [… adapted] |
| *IG dimension: IG general, legal, …* | If an entry covers several IG dimensions, it is *indicated in italics* in abbreviated form. |
| ***Internal: … | "***Internal: …" indicates internal notes.<br>(only used during the preparatory stages of the IGG project) |

For the sake of clarity and user-friendliness, lexicographical symbols (such as different kinds of parentheses, abbreviated indications etc.) are kept to a minimum and used consistently.

## 2.2 TERM USAGE AND RELATIONS

Although certain 'prescriptive' aspects are applied in the IGG, it cannot be considered as 'normative' in the sense that it is approved by a domain authority with normative powers, such as a standardising organization. Nevertheless, one can and sometimes has to make a difference between preferred term, admitted term and deprecated term, if there is more than one term or name for a given concept. In rare cases, even an abbreviated form may have a status different from its full form.

In the context of the IG glossary, a usage status of a term can be one of three types (preferred term, admitted term and deprecated term) in analogy to the definitions in ISO 10241-1:2011 adapted here for the purpose of the IGG:

**1. Preferred term**

rated as the primary term or name for a given concept. There can be more than one preferred term or name. If there is only one term representing the concept, this term is automatically preferred. By analogy, 'preferred' can apply also to abbreviated forms. Preferred terms or names are written in bold letters.

→ Example 1: full form and abbreviated form both being preferred terms
**Internet service provider**
**ISP**

→ Example 2: abbreviated from and full form both being preferred terms (but one of them may be more often used)
**IPv4**
**Internet Protocol version 4**

**2. Admitted term**

a synonymous term or name for a preferred term, but not rated as a preferred term. There can be more than one admitted term. By analogy, 'admitted' can apply also to abbreviated forms. Admitted terms or names are written in regular letters.

→ Example 1: two preferred terms and an admitted term
**eSignature**
**electronic signature**
digital signature

→ Example 2: abbreviated form as preferred term and full form as admitted term
**IP address**
Internet Protocol address

### 3. Deprecated term

a synonymous term or name for a preferred term, but not rated as preferred term or admitted term. There can be more than one deprecated term. By analogy, 'deprecated' can apply also to abbreviated forms. In order to mark deprecated terms, they are preceded by "NOT:"

→ Example: preferred term and deprecated term
**Internet operator**

↳ NOT: carrier
*Explanation: Both terms are widely used as standardized terms (have definitions/ standardized meaning). However, the term "carrier" is used too widely in different fields, such as transport and ICT, and could be misleading in the IG context. For the sake of clarity, the term "Internet operator" is the preferred term to be used within IG community instead of the term "carrier".*

### 4. Related term

In the case that a conceptually near term is a 'quasi-synonym' – i.e. not a fully synonymous term – or a closely related term, it is marked by a preceding "RT:" (i.e. 'related term'; similar to the use in thesaurus development).

→ Example: preferred term and related term
**Internet Protocol suite**

RT: TCP/IP
*Explanation: Why RT? TCP and IP are the fundamental protocols of the Internet Protocol suite, which also contains other protocols. On the other hand, TCP/IP is often used as synonym for the Internet Protocol suite.*

## 2.3 ORTHOGRAPHIC AND OTHER WRITING CONVENTIONS

For some terms, names or words there may be different orthographic variants – not only regional variants, such as British English vs. American English – for which decisions had to be taken in order not to inflate the number of synonyms:

→ Example: eCommerce

↳ NOT: e-Commerce, e-commerce, E-commerce etc.

In the IGG these eTerms are written without hyphen, small e and initial capital letter for the application area it refers to, such as "eCommerce".

↳ *Explanation: Hyphens and other syntactic signs in terms are used as in their most frequently occurring form in texts. Orthographic and other writing conventions are bothering in term extraction across many different documents (especially when including different sorts of text). Each variant has to be searched separately – or settings (such as case-sensitivity) have to be modified with each search.*

In some cases, it is difficult to distinguish between a general term and a name, such as in the case of internet vs. (the) Internet, or Internet Protocol (IP) vs. some other internet protocols. Therefore, terms appearing as entry terms in an IGG entry are consequently written with lower case letter at the beginning, whereas names are capitalized.

Needless to mention, the above-mentioned phenomena pose serious obstacles for obtaining satisfactory results by using term extraction tools. (See part 3)

## 2.4 DESCRIPTION OR DEFINITION?

Although it was intended in the beginning to use definitions in the strict sense, the idea was abandoned in the course of the work for a number of reasons:

- Strict and concise definitions are obligatory in highly normative and systematic terminology approaches which tend to be geared towards achieving complete terminologies. But the IGG aims at a selection of major terms and names used in IGG and, therefore;
  - → is not under the requirement to be complete;
  - → descriptions may sometimes have to contain redundant elements or be less 'technical' in some wordings;

- The requirement of "user-friendly" definitions in combination with a selective vocabulary inevitably leads to 'descriptions'.

However, this does not imply that descriptions are less correct than definitions. On the contrary, it was a particular challenge to find the right wording for each description, so that it makes sense in the context of other descriptions used in the IGG, in view of the target audience.

Another challenge was to connect the descriptions to each other through cross-referenced terms, turning the whole IGG into a coherent context. User-friendliness in this connection means that these terms constitute a supportive information for the user in the related IGG entries. In other tools of reference, the user is often confronted with conflicting or even contradicting information

## 2.5 MACROSTRUCTURE OF THE IGG

In addition to the title pages, the IGG comprises the following sections:

- Background
- Table of contents
- Table of terms
- The glossary sections
- Annex 1: List of references used for the preparation of the glossary of IG terms
- Annex 2: Interrelationship of IG dimensions
- Annex 3: Alphabetical index of terms and names

For the macrostructure of the glossary entries, the IG dimensions, largely following the widely accepted IG dimensions as sketched by Kurbalija (2012), suggested themselves as major differentiation aspects for structuring the IGG. (See Figure 6) Therefore, the glossary part is sub-divided into 5 sections in line with the IG dimensions, each preceded by a short introduction. They are followed by Section VI comprising the types of stakeholders, while Section VII contains proper names of existing organizations, forums, networks, groups, conferences, regulations and legal instruments:

**I.** Internet governance in general
**II.** Infrastructure and standardization
**III.** Economic dimension
**IV.** Legal dimension
**V.** Development and socio-cultural dimension
**VI.** Stakeholders
**VII.** Organizations, forums, networks, groups, conferences, regulations and legal instruments

Sections I to V – while largely following the widely accepted IG dimensions as sketched by Kurbalija (2012) – were modified for IGG purposes to the extent that

- a 'meta-dimension' "Internet governance general" was introduced,
- socio-economic development and the sociocultural line were merged into one dimension,
- the dimension of stakeholder categories was added.

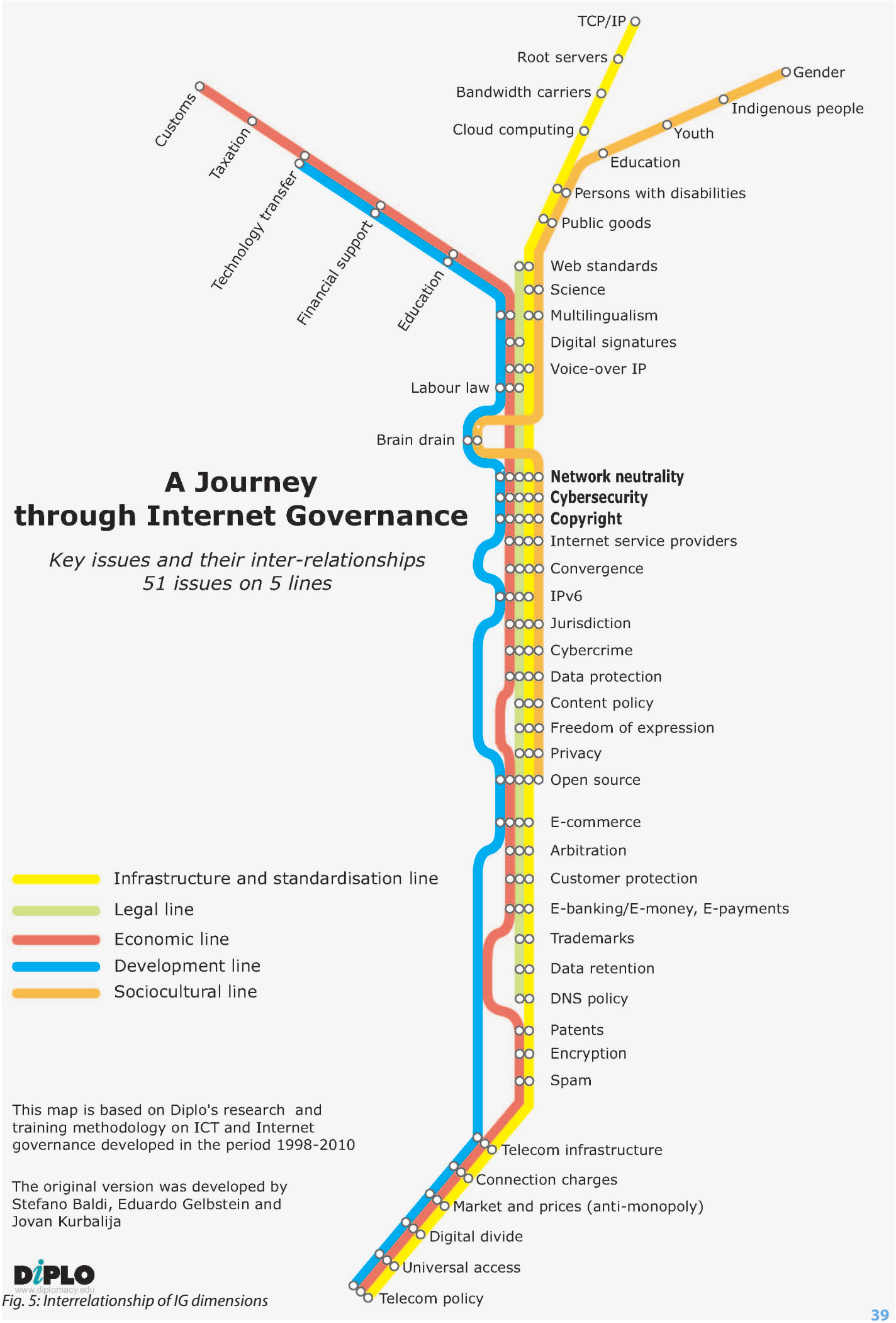


*Fig. 5: Interrelationship of IG dimensions*

# OBSERVATIONS ABOUT THE USE OF THE TERM IDENTIFICATION AND EXTRACTION TOOL PROTERM

The application of computational linguistics software for the purpose of compiling the IGG turned out to be a challenge. Most of the tools have been developed for different purposes and approaches.

## 3.1 TESTING SEVERAL TERM EXTRACTION TOOLS

Right at the beginning of the project, several term identification and extraction tools were tested. Most of the easily available tools had:

- limitations with respect to volume of text to be processed (e.g. 5 MB),
- constraints concerning the formats of the texts to be processed,
- many of these tools are programmed for bilingual extraction or ask users to feed translation memories before any action can be carried out,
- flaws in identifying term candidates e.g.
  - with respect to taking line by line of the documents as basic units for term identification (thus for instance cutting a multiword term or hyphenated term into two parts at the end of the line and taking the parts as different terms),
  - concerning the inability to distinguish between names of organizations and terms.

In some cases, the documents had to be de-formatted first (e.g. taking out title pages, graphs etc.) before they could be loaded into the system for processing.

When applying term extraction tools, it became clear that frequency of occurrence is not an ideal indicator for the importance of a given term candidate:

→ In some cases, there were only one or very few occurrences which proved highly significant.

→ In other cases, there were lots of occurrences which proved more or less to be insignificant.

→ Significance of terms/contexts often had to be checked against additional sources.

In addition, many terms occurred in two or more variants (e-signature, E-signature, eSignature etc.), in one case, two full forms had the same abbreviated form (CDN), many terms had synonyms or quasi-synonyms. In most cases, it was comparatively easy to identify multi-word terms, whereas some mono-word terms, such as "openness" posed difficulties. This applies to manual term extraction as well as to automatic term extraction – however, in manual term extraction the human brain is more effective in sorting out relevant from irrelevant data.

The term and name identification tool ProTerm proved to have much less limitations and constraints than other tools tested. In addition, several corpora from the same selection of documents in ProTerm could be used side-by-side – e.g. for different search strategies.

Therefore, ProTerm was chosen (and adapted) and persons trained to professionally use it.

## 3.2 EXPERIMENTING WITH PROTERM

When experimenting with ProTerm, it was soon found out that ProTerm identified too many term candidates, because:

- the collection of documents not only covered IG terminology, but also other fields,
- IG terminology (incl. neighbouring fields) alone would – if going into very specific detail – comprise more than 1000 terms.

Therefore, it was decided to use at the beginning of the IGG project the most "explicit" documents for manual term extraction and recording of contexts in the preparation of formulating the descriptions with computer assistance later.

## 3.3 IDENTIFYING PERTINENT TERMS AND NAMES

The list of terms and names identified were checked and evaluated. All identified terms at that stage were considered pertinent, but about half of them were de-selected, as :

- the experts had other priorities from those reflected in the documents, (which is natural, not to mention that there is a lot of content duplication as well as different emphasis in the documents),
- the authors (or issuing organizations) of the documents probably had quite different professional backgrounds from those asked to make a selection of the most important IG terms,
- some documents were reflections, agendas or confrontations that were not adequate for a description.

However, new IG terms are emerging virtually by the day. That is why some experts added a few new entries, while others wanted to delete entries.

## 3.4 EFFICIENT USE OF PROTERM

Based on the consolidated list of term candidates for the IGG, ProTerm could be applied in a targeted way, showing good results for the already identified terms/names and their contexts. The system showed the frequency of their occurrences (sometimes with different results depending on orthographical variants; others with different word combinations; etc.). It also permitted to check each term occurrence in the tool or in the original document. This possibility contributed to clarify/harmonise the explanations and filter the differences and similarities in the respective field or subfield.

In any case, the results of applying ProTerm at that stage were categorically better than using ProTerm for term identification without preselected terms. In this connection, some interesting observations could be made with respect to the frequency of occurrences of terms and names in the collection of documents:

- A high frequency of occurrences may be an indication of the high relevance of a term or name – or may be not in other cases;
- A low frequency may not necessarily be an indication of a low relevance of a term or name – sometimes on the contrary;
- In some cases, a term was not found at all in the initial set of documents, although considered highly relevant – possibly because of its novel nature.

ProTerm evidenced that the contexts for each term or name were often semantically "poor" so that the descriptions had to be enhanced by combining several contexts or using additional sources.

By means of ProTerm it was possible to look at the context not only during the process of term identification (when it is displayed in plain text format), but also check the contexts in the documents residing in the tool in their original version with the original layout and in combination of non-verbal representations. Therefore, it is necessary to look into the original document many times in order to understand some contexts.

Even though the tool is very useful, the need to compare and verify contexts individually for each term is time-consuming. It is indispensable to go term by term to identify the contexts that provide significant information, needed to formulate the description of the term. Often useful information was found in documents that were recommended to be deleted from the main list. Therefore, it is good to be able to use several corpora (each composed of different sets of documents) in the same tool.

At the beginning, all the documents were uploaded and three databases were prepared:

- The first one contained all documents.
- The second contained only the selection of documents that were colour-coded according to their relevance.
- The third one was constituted only for the most relevant documents together with a 'stop word' list with the terms identified in the literature during the 5th step of the workflow; thus, additional terms could be identified as core terms and added to the list of term candidates for the IGG.

After that, it was necessary to clean the terms candidates from all kinds of 'noise' in order to obtain precise results. Otherwise, the tool would identify more term candidates because they are accompanied by marks (such as syntactic signs) or other elements/features that show them as different terms. The deletion/suppression of these marks/features assures more precision with respect to the frequency of occurrence in different documents, the completeness of identifying term candidates and a higher chance of detecting significant contexts.

This can be exemplified by the following screen shot:

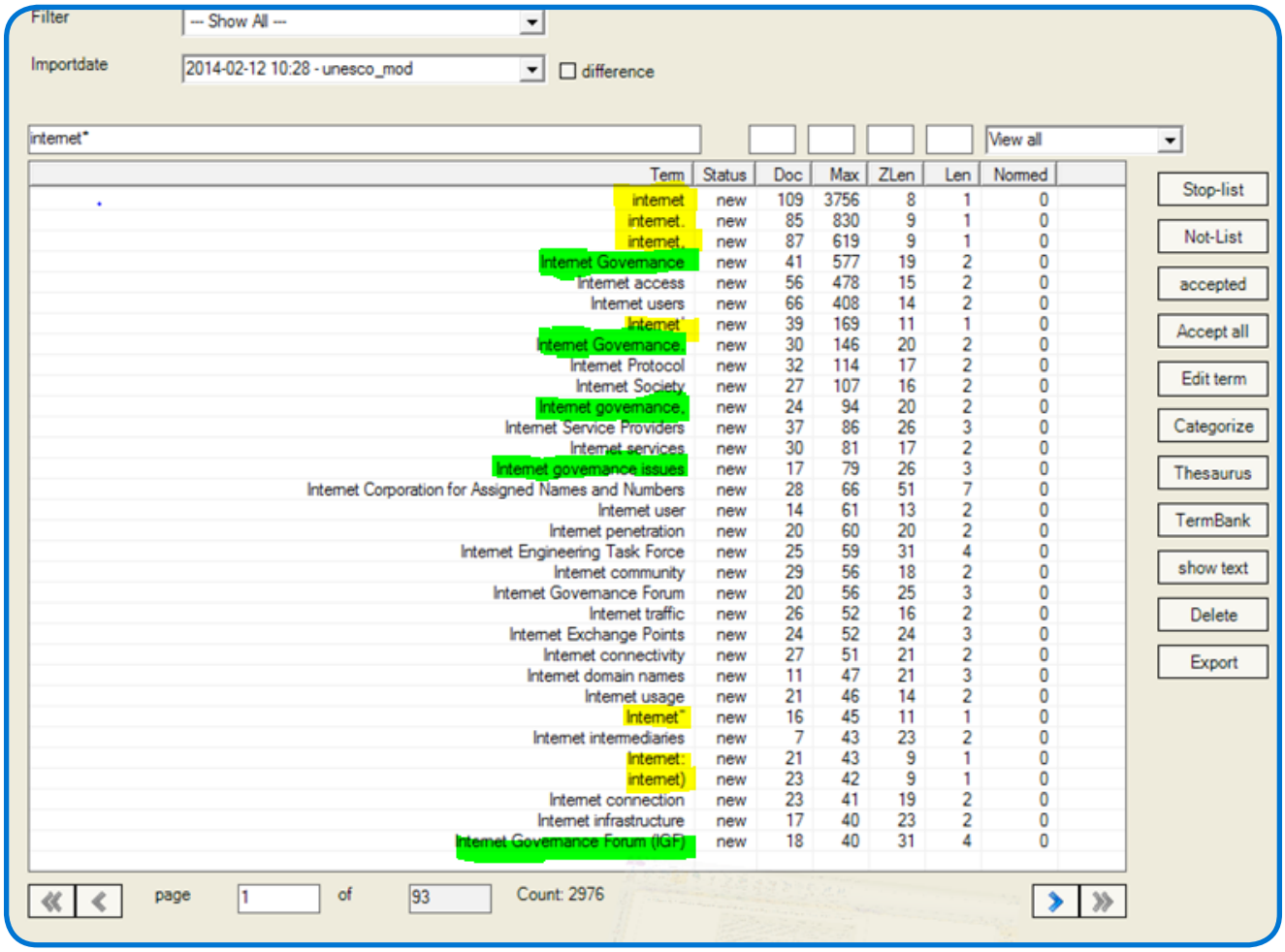


*Fig. 6: ProTerm screen for clustering terms containing "Internet"*

When looking for 'internet' you find 7 forms of occurrences (internet – internet, – internet. – Internet* – Internet: – Internet) on the same page and in different quantities of documents. Each form should be edited in order to harmonise the term layout e.g. by separating punctuation marks etc. from the term. In this example we can see internet with period, comma, apostrophe, asterisk, colon, and parenthesis. In the case of Internet governance, three forms of occurrences were "different" because of punctuation marks, but the other two were new terms closely related to the base term Internet governance.

ProTerm allows searching for terms and names by looking for a 'word' in co-occurrence with other words on the right or on the left to a given 'word'. This facilitates the detection of compound terms and colocations.

The possibility to open the three ProTerm databases established for the term/name identification, simultaneously or the main database several times, permits to contrast results, to verify possible combinations and to find information scattered in different documents. This contrasting of contexts

is extremely helpful to consolidate the descriptions. At the same time, duplication of (similar or contradicting) information across several documents is evidenced.

One of the advantages of ProTerm is that it is possible to work with monolingual or bilingual texts depending on the needs. Text files may be in different formats. It also permits to carry out several activities in the same system environment. This is different from other tools that specially work with bilingual or multilingual texts for other purposes.

# GENERIC APPLICCABILITY OF THE METHODOLOGY

Internet governance is a highly dynamic and multi-facetted field where terminology will continue to emerge. This applies in a different way also to other fields, such as the management of enterprises where management strategies and measures have to be continuously – sometimes immediately and radically – adapted to the requirement of the markets. For practical management, too, a 'user-friendly' terminology was preferred to an academic-scientific one.

Updating and expanding a glossary, such as the IGG, is an essential requirement in order to keep the data valid and reusable over time for the sake of content sustainability. For the purpose of updating and expanding such glossaries, there are other tools available which by themselves or in combination with a tool like ProTerm can support this updating and expansion process and render it highly efficient.

In the course of updating and expansion it is important to keep record – i.e. the history – of all deleted entries as well as modifications in maintained entries. Deleted entries may need to become reinstalled in the course of development. The history of modifications saves a lot of time spent on re-discussing questions already having been solved in the past. In the IGG this happened with the term 'openness'.

In the case of the IGG, the final face-to-face meeting proved to be greatly significant taking into consideration that the primary focus was on the Arabic version. It is always advisable, when preparing such glossaries, to bear in mind future needs for translation into other languages. In this connection, the approaches of controlled language or simplified language may be useful. This applies particularly to cases of 'transcreation' into really foreign languages when issues of different scripts, writing conventions and cultural requirements have to be taken into account. For the IGG controlled language approaches were applied.

This makes the IGG methodology applicable to many other glossary endeavours of a similar nature. The more diversified the requirements for a glossary based on terminological principles are, the more the methodology used for the development of the IGG will have to be adapted to the intended purpose. To give an example, the IGG has been prepared (for the time being in English and Arabic) in such a way that the final result can easily be re-used for other purposes and updated or upgraded in various directions by:

- adding further languages
- using conventional or advanced technical platforms

The above directions are closely intertwined.

The IGG, as it is, contains so much useful data in user-friendly presentation, that it could be used as an educational resource (e.g. as open educational resource – OER). As soon as other languages and adaptations for other purposes are aimed at, cooperative/participatory methods and tools should be considered, in order to attain a high level of content sharing.

United Nations
Educational, Scientific and
Cultural Organization

Information for All
Programme


IGG