Deliverable D6.6.1

QTLaunchPad Workshop: Quality Metrics for Human and Machine Translation

Author(s): Stephen Doherty, Federico Gaspari, Josef van Genabith, Declan Groves, Ankit Srivastava (DCU)

Dissemination Level: Public

Date: 24.04.2013

SEVENTH FRAMEWORK PROGRAMME

D6.6.1: QTLaunchPad Workshop: Quality Metrics for Human and Machine Translation

Grant agreement no.	296347
Project acronym	QTLaunchPad
Project full title	Preparation and Launch of a Large-scale Action for Quality
Funding scheme	Coordination and Support Action
Coordinator	Prof. Hans Uszkoreit (DFKI)
Start date, duration	1 July 2012, 24 months
Distribution	Public (Internal Version)
Contractual date of delivery	March 2013 - Extended
Actual date of delivery	April 2013
Deliverable number	D6.6.1
Deliverable title	QTLaunchPad Workshop on Quality Metrics for Human and Machine
Type	Report
Status and version	Internal Draft for Review
Number of pages	28
Contributing partners	DCU, DFKI, USFD
WP leader	DCU
Task leader	DCU
Authors	Stephen Doherty, Federico Gaspari, Josef van Genabith, Declan
EC project officer	Kimmo Rossi
The partners in	Deutsches Forschungszentrum für Künstliche Intelligenz (DFKI),
	Dublin City University (DCU), Ireland
	Institute for Language and Speech Processing, R.C. “Athena”
	The University of Sheffield (USFD), United Kingdom

D6.6.1: QTLaunchPad Workshop: Quality Metrics for Human and Machine Translation

For copies of reports, updates on project activities and other QTLaunchPad-related information, contact:

DFKI GmbH

QTLaunchPad

Dr. Aljoscha Burchardt	aljoscha.burchardt@dfki.de
Alt-Moabit 91c	Phone: +49 (30) 23895-1838
10559 Berlin, Germany	Fax: +49 (30) 23895-1810

Copies of reports and other material can also be accessed via http://www.qt21.eu/launchpad

Contents

1 Introduction

The aim of this report is to capture the discussions and feedback from two public workshops on the QTLaunchPad quality metrics on human and machine translation. Following an introduction to the multidimensional quality metrics (MQM) being developed in the project and a description of the workshops' format, this report:

1. provides specific feedback and discussion points for the MQM framework (Section 2);
2. highlights key commonalities (Section 2);
3. recommends future steps informed by the above (Section 3).

1. 1 Multidimensional Quality Metrics

As a core aspect of the QTLaunchPad project, the multidimensional quality metrics framework (MQM) is a means of moving beyond the current shortcomings of existing translation quality assessment. Quality assessment (QA) is an important task in the translation workflow, especially in the context of machine translation. Traditional methods are typically very subjective and involve assessors counting errors and their severity. This approach has led to the formalisation of metrics for assigning errors to different types (e.g. incorrect spelling, incorrect terminology, wrong verb agreement) and counting their prevalence and severity in a random sample of translated content. This approach developed into specifications such as the LISA QA Model and SAE J2450, however, these models have not been updated consistently and have shortcomings in terms of validity and robustness.

The prevalence of the use of MT in translation and localization workshops also presents a challenge for evaluation and QA processes due to the unique nature of MT output, and the extensiveness of human intervention, e.g. use of raw MT output, human post-edited MT. The "one-size-fits-all" approach of existing models simply cannot meet the needs and expectations of a rapidly diversifying translation industry.

Following a systematic review of existing models and extensive public consultation, the QTLaunchPad project is developing a new framework for assessing quality based on the following principles:

- Adaptability - metrics must be adaptable to specific project types as projects, users, requirements, and scenarios are unique. This allows for metrics to be tunable and sensitive to each application;

- Granularity - metrics must allow for customizable degrees of granularity, from extremely coarse to extremely fine, depending on the use case, e.g. evaluation of gist translation vis-á-vis a detailed analysis to isolate errors;
- Comparability - results must be comparable and meaningful across jobs, projects, users, and domains of application;
- Suitability - metrics must be suitable for both human and machine translation and combinations thereof in addition to all technology and production profiles and users;
- Standardised - metrics must develop upon existing standards and established best practice in order to optimise this progress while supporting interoperability, proprietary methods, and customisable workflow integration;
- Fairness - existing metrics conflate errors in the source, errors in the target, and actual translation errors. This results in problems identifying the cause of the problem, where translators are often blamed. QA should be fair and recognise the work of translators, especially when they add value in this way.

A more extensive and practical description, and further information on implementation, is available from the project website[1], and in the publicly available deliverable D1.1.2: Multidimensional Quality Metrics.

1.2 MQM Workshop and Related Events

As a means of information gathering from relevant stakeholder groups, a survey was conducted in late 2013 (whose results are documented in project deliverable D6.5.1: Report on Requirements Gathering from Relevant Stakeholder Groups). Of 438 respondents across approximately 40 countries, the survey examined a cross-section of the main stakeholder groups targeted by the QTLaunchPad project: translators and LSPs, large-scale public users of language technology (LT), providers of LT, and corporate users.

The findings of the survey highlighted current trends and best practices, and were particularly insightful regarding the utilisation of language and translation technologies in the context of quality assessment and MT post-editing. The survey also identified shortcomings, for instance, in the absence of comparable and robust QA methods and in the quality thresholds of current approaches to MT.

[1] http://www.qt21.eu/launchpad/content/multidimensional-quality-metrics

Informed by these findings (D6.5.1), several general topics of importance were identified and appear below as ranked by workshop participants prior to the event[2]:

- the need to move beyond current approaches to evaluation;
- the availability of high-quality corpora, datasets, systems, etc.;
- funding for language and translation technologies;
- the need for specialised knowledge and expertise;
- the quality of current corpora, datasets, etc.;
- uncertainty about the future of MT.

The project consortium identified co-location with the MultilingualWeb workshop in Rome (March 14th, 2013) to be the ideal first venue for the QTLaunchPad MQM framework presentation due to its location, resource cost, and audience (attendance of 150). The W3C MultilingualWeb[3] workshops are free and open to the public and concern standardisation and best practice for multilingual and multicultural web-based information, e.g. localisation, code standards. In sponsorship of the event, QTLaunchPad received e-mail access via the organiser to the attendees, a full-page promotional piece for the project, and space for exhibition in the main room of the event (pull-up poster and flyers).

The final agenda for the Rome workshop can be found in Appendix A. The MQM workshop took place in the morning of the full-day session, where the afternoon was assigned to the research innovation application scenarios, RIAS, as detailed in D6.7.1. Despite considerable last-minute cancellations both the W3C and QTLP workshops have encountered, each of the aforementioned stakeholder groups were well represented at the workshop: 7 LSPs, 2 LT providers, 2 corporate users, and 3 large-scale public users. The structure of the MQM workshop is outlined below and corresponds to the structure of feedback in Section 2 of this report:

- Overview of Metrics:
    - Principles
    - Issue Types
    - Dimensions
- Using Dimensions to Build Task-Relevant Metrics
- Demonstration of MQM Tool

[2] http://docs.google.com/spreadsheet/ccc?key=0AotdkIT3g3R8dEZsa2pVVElhQk5QT3FBZkRlT3Vkc1E&usp=sharing

[3] http://www.multilingualweb.eu/

- Closing Discussion

QTLaunchPad also ran several events at the Globalization and Localization Association (GALA) conference in Miami[4] (March 17th-20th with an attendance of approximately 250 delegates with QTLaunchPad flyers in each bag) supported by the project's sub-contractor GALA. GALA is the world's largest association for the language industry, including translation and localization. It is a non-profit in nature and provides resources to thousands of its members. The QTLaunchPad presence consisted of:

1. a 45-minute talk on MQM as part of the main conference programme (43 attendees in three-way parallel slot);
2. an interactive exhibition open through the conference;
3. a roundtable discussion on MQM (two 90-minute slots).

The exhibition space made use of the same marketing materials as the Rome event: flyers detailing the MQM, and a pull-up banner. The combination of events allowed project representatives many opportunities to present QTLaunchPad, the RIAS, and metrics to this audience, which resulted in feedback and contact points for many aspects of the project work, including feedback on RIAS as outlined in this report, the multidimensional quality metrics (D6.6.1), and future directions (D6.8.1 and D6.8.2).

Promotional material for these events can be found on the QTLaunchPad website[5], the MultilingualWeb website[6], the workshop flyer[7], e-mail and social media announcements[8], and on the GALA website[9]. In addition to these channels, the workshop was advertised via e-mail lists from project partners, including CNGL, MT-List, and social media (e.g. Twitter[10] and LinkedIn[11]).

[4] http://www.gala-global.org/conference/

[5] http://www.qt21.eu/launchpad/content/workshop-research-innovation-application-scenarios-%E2%80%93-rome-march-14th-2013

[6] http://www.multilingualweb.eu/documents/rome-workshop/rome-program

[7] See Appendix B

[8] See Appendix C

[9] http://www.gala-global.org/conference/qtlaunchpad-showcase

[10] http://twitter.com/qtlaunchpad

[11] http://www.linkedin.com/groups?gid=4807518

Prior to the above events, QTLaunchPad also ran a webinar[12] via GALA to showcase the MQM framework (February 21st, 2013). Following the presentation there was time for questions and discussion from the 75 participants (out of a total of 162 registered), and additional follow-ups via the project's LinkedIn page and opt-in e-mail list[13].

[12] A recording of the webinar can be found here: http://www.gala-global.org/recordings-past-webinars#GSI and a PDF of the content here: http://www.gala-global.org/files/21Feb2013_QTLaunchpad_PPT.pdf

[13] http://www.dfki.de/mailman/cgi-bin/listinfo/qtlp-news%20

2 Workshop Feedback

The feedback gathered at the events described has considerably helped shaping the ongoing project developments and has directly been incorporated into the respective Deliverables. This section summarizes main points from the feedback and discussions of the above workshop and related events into the categories of: metrics and applications.

2.1 Metrics

- While it's not always possible, it's important to know the contexts in which metrics are used. This includes knowledge of: users, domains, formats, tools, resources, etc.
- Taking source text quality in a translation metric into account is a positive development, but may not be immediately possible in existing workflows.
- The range of issue types in the MQM needs to be practical and contain customisable degrees of granularity. It is perhaps not possible to find a balance that suits all users, so some level of customisation is necessary, e.g. Canadian government's SEPT error categories contains approximately 700 aspects, while very comprehensive it is not applicable in industrial application.
- There is also a need to look at evaluation processes, not just at error metrics. TAUS has a Dashboard that gives options for job types, end use, and tools, etc., but it fails to address how the evaluation data are used and if they are meaningful and effective or not.
- Further to this, the issue of comparability across jobs, projects, and evaluation paradigms is burdensome. Metrics such as BLEU have become the de facto standard in research, yet may not be at all meaningful for translators or buyers - MQM must work with the existing evaluation landscape.
- Further attention should be given to the nature of error categories rather than errors themselves; this may be fruitful in terms of pre-processing, standardisation, and overall resource saving.
- There will need to be well-documented and easily accessible and understandable content for the basis of the MQM, its usage, and its value over existing approaches.
- Sensitivity to the different types of errors introduced by human translation, machine translation, and combined approaches.
- There should be a balance of preventative and reactive strategies for QA - human in the loop, rather than effort wasted, e.g. post-edited content and evaluation data to be used to improve MT processes - MQM needs to include such a balance.

- Finally, the validation of the MQM is critical. To learn from existing models, there needs to be strict reliability and validity testing, etc.

2.2 Applications

- It is well recognised that current models do not meet industry needs. It is apparent to most stakeholders that there is a need for updating existing approaches to evaluation and QA rather than following the status quo in a haphazard, internally-focused, and reactive way.
- Translators are typically not consulted in the development of metrics and their usage. Inclusion of these groups will be valuable, especially in the proposition of the MQM that translations are penalised for correction errors in the source.
- Evaluators are busy and sometimes even ‘lazy’. Despite an extensive list of errors, the same small number of metrics are used from a drop-down menu.
- The ambiguity of certain tools or types (of errors) can be a source of the previous point, but more generally, can lead to poor evaluation results that can vary greatly from person to person. To combat this, ambiguity should be reduced by using clearly defined types and procedures with cooperation between the tool and the users. (to suit their needs).
- It is also ambiguous what is meant by compatibility in MQM. Such ubiquitous compatibility with existing models is a claim that may be difficult to accomplish in practice.
- In terms of formats: what will MQM be compatible with, what options are there with regard to input/process/output formats and encoding?
- The evaluation/QA data from MQM should also be meaningful, customisable, and allow for different degrees of granularity, otherwise the best features of the metrics may be lost or ignored post-evaluation.
- Alongside MQM, the approach of avoiding errors in the first place should be pushed - preventative rather than reactive steps. Thus, standardisation and pre-processing are necessary components to high-quality translation results. Translation quality is linked directly to pricing - higher quality equates to higher compensation. It is therefore important to establish thresholds, just like production workflows.
- In terms of usage, how can issues with representative sampling be addressed in MQM? Currently, there is a need to sample and assess in QA models, while AEMs like BLEU can assess the whole document/system.

- There needs to be a clear way for users of MQM and the tool to give feedback and be more interactive with the project - mailing lists and one-way communications are not sufficient.
- Open field testing will allow for the inclusion of as many viewpoints as possible for different users in a variety of scenarios.

3 Conclusion

3.1 Summary

The feedback QTLP has gathered was by and large very positive, supporting the work that has been done in the first phase of the project. The project was also confirmed in choosing an early communication strategy exposing even ongoing work to professional criticism.

In summary, the focus of stakeholders and users is apparent throughout the workshop discussion and feedback. Stakeholder/user inclusion and buy-in are pivotal in the uptake of new tools such as the MQM. Therefore, the entry level for its use must be low, with support and accessible materials from the onset. The value of such adoption must be clear and quantifiable for it to be successful. The more specific feedback relating to the metrics will be addressed in the update to the first version of the MQM; the second, revised version will be available in project deliverable D1.1.2.

Further to the acquisition of this valuable discussion and feedback outlined in the previous section, the workshops resulted in quantifiable gains in terms of QTLaunchPad's public exposure (traffic to the project homepage, www.qt21.eu, increased from 302 unique visits in the month of February to 524 in March, and 168 in April[14]), awareness raising of RIAS topics, and membership to the project in terms of social media (e.g. increase to 71 members on LinkedIn[15]), mailing list (increase to 343 new opt-in members), and individual contacts with stakeholders.

3.2 Future Directions

In addition to the above, future directions for the MQM were invited where the following points were discussed:

- The project's critical mass and development of the MQM represent sufficient resources for large-scale field testing and further refinement with community/industry input.
- There needs to be an ongoing collaborative process where quantifiable value is evident for all parties, especially from industry buy-in viewpoint. In the context of MQM, this may result in collaborative efforts to test the framework in a variety of real-world contexts where feedback from users is paramount.

[14] As of April 15th - to be updated at month end.

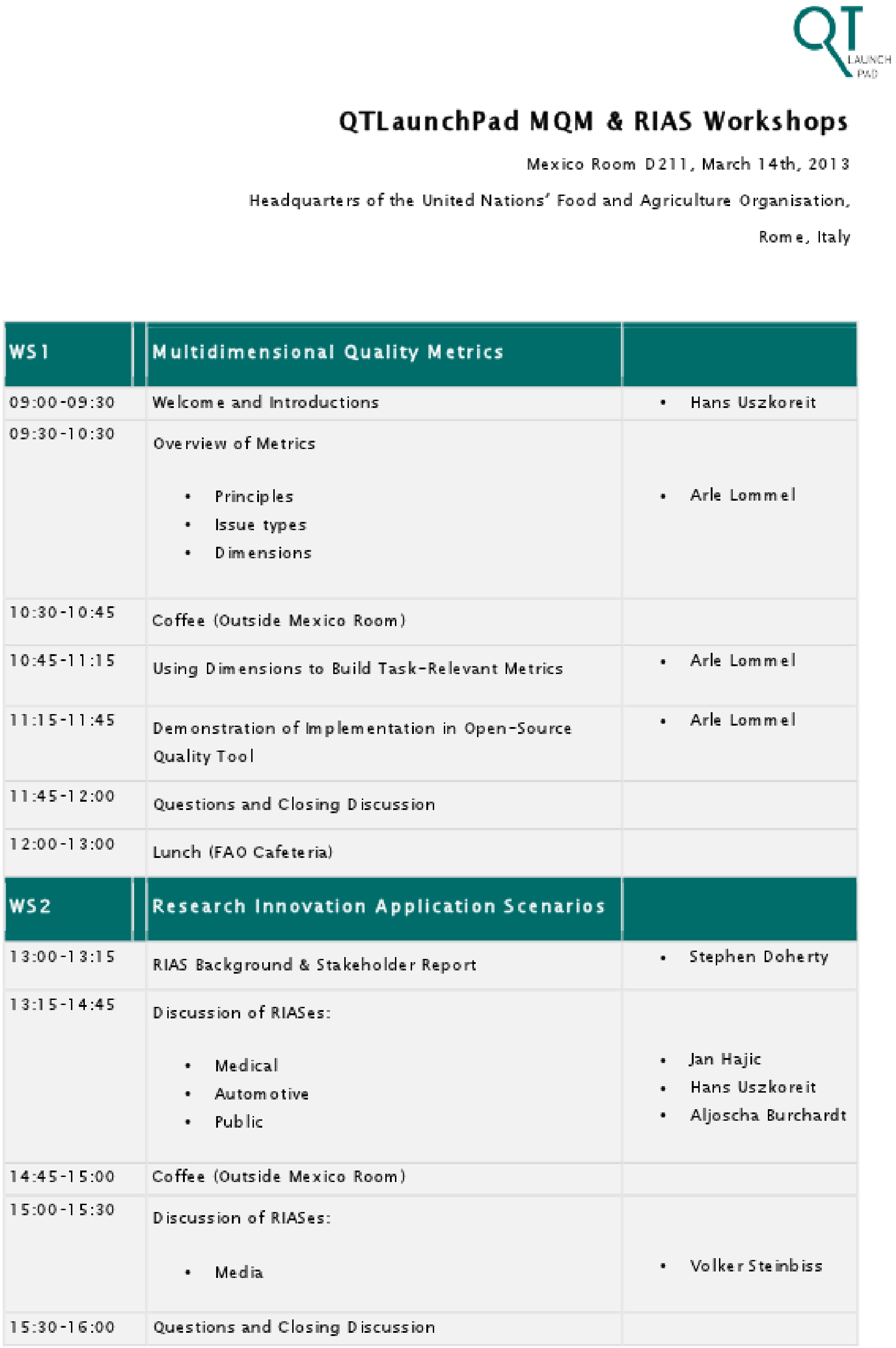
[15] For comparison, the META LinkedIn group has 356 members.

- The roll-out of the MQM platform, translate5[16], needs to be coordinate effectively and user-centric to ensure the greatest uptake and effective incorporation of feedback. This could come in the form of usability, feedback sidebar, two-way communication between users and QTLaunchPad, education and educational materials, incentivisation for users to give feedback on errors, crashes, etc.
- Feedback should be documented and the improvement of the MQM with this value input should be made clear to support ownership and adoption of MQM by the user community.

[16] http://www.translate5.net/

Appendices

Appendix A - Rome Workshop Agenda



QTLaunchPad MQM & RIAS Workshops

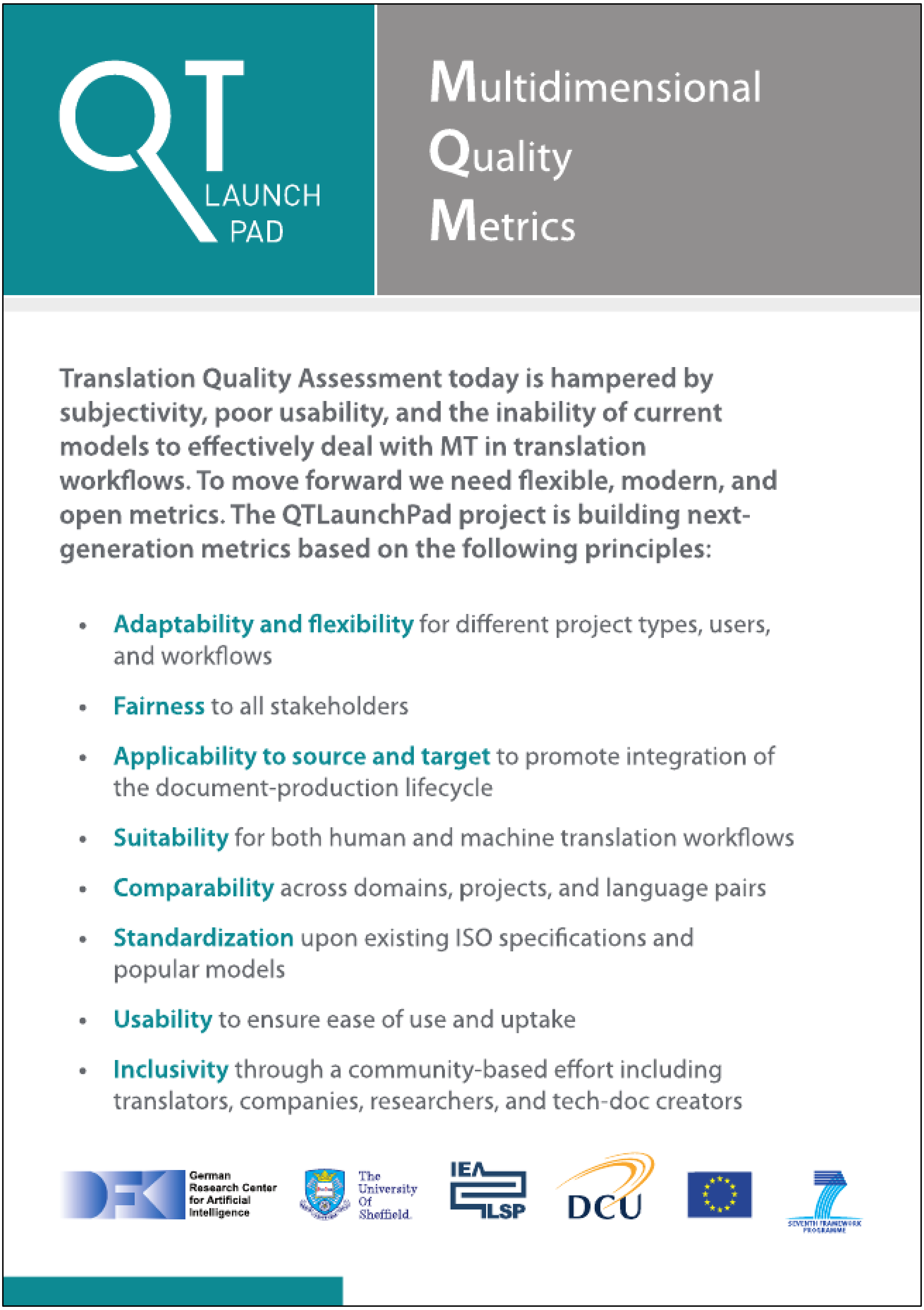
Mexico Room D211, March 14th, 2013

Headquarters of the United Nations' Food and Agriculture Organisation,

Rome, Italy

WS1	Multidimensional Quality Metrics	
09:00-09:30	Welcome and Introductions	• Hans Uszkoreit
09:30-10:30	Overview of Metrics • Principles • Issue types • Dimensions	• Arle Lommel
10:30-10:45	Coffee (Outside Mexico Room)	
10:45-11:15	Using Dimensions to Build Task-Relevant Metrics	• Arle Lommel
11:15-11:45	Demonstration of Implementation in Open-Source Quality Tool	• Arle Lommel
11:45-12:00	Questions and Closing Discussion	
12:00-13:00	Lunch (FAO Cafeteria)	
WS2	**Research Innovation Application Scenarios**	
13:00-13:15	RIAS Background & Stakeholder Report	• Stephen Doherty
13:15-14:45	Discussion of RIASes: • Medical • Automotive • Public	• Jan Hajic • Hans Uszkoreit • Aljoscha Burchardt
14:45-15:00	Coffee (Outside Mexico Room)	
15:00-15:30	Discussion of RIASes: • Media	• Volker Steinbiss
15:30-16:00	Questions and Closing Discussion	

Appendix B - Workshop A5 Flyers



QT LAUNCH PAD

Multidimensional Quality Metrics

Translation Quality Assessment today is hampered by subjectivity, poor usability, and the inability of current models to effectively deal with MT in translation workflows. To move forward we need flexible, modern, and open metrics. The QTLaunchPad project is building next-generation metrics based on the following principles:

- **Adaptability and flexibility** for different project types, users, and workflows
- **Fairness** to all stakeholders
- **Applicability to source and target** to promote integration of the document-production lifecycle
- **Suitability** for both human and machine translation workflows
- **Comparability** across domains, projects, and language pairs
- **Standardization** upon existing ISO specifications and popular models
- **Usability** to ensure ease of use and uptake
- **Inclusivity** through a community-based effort including translators, companies, researchers, and tech-doc creators

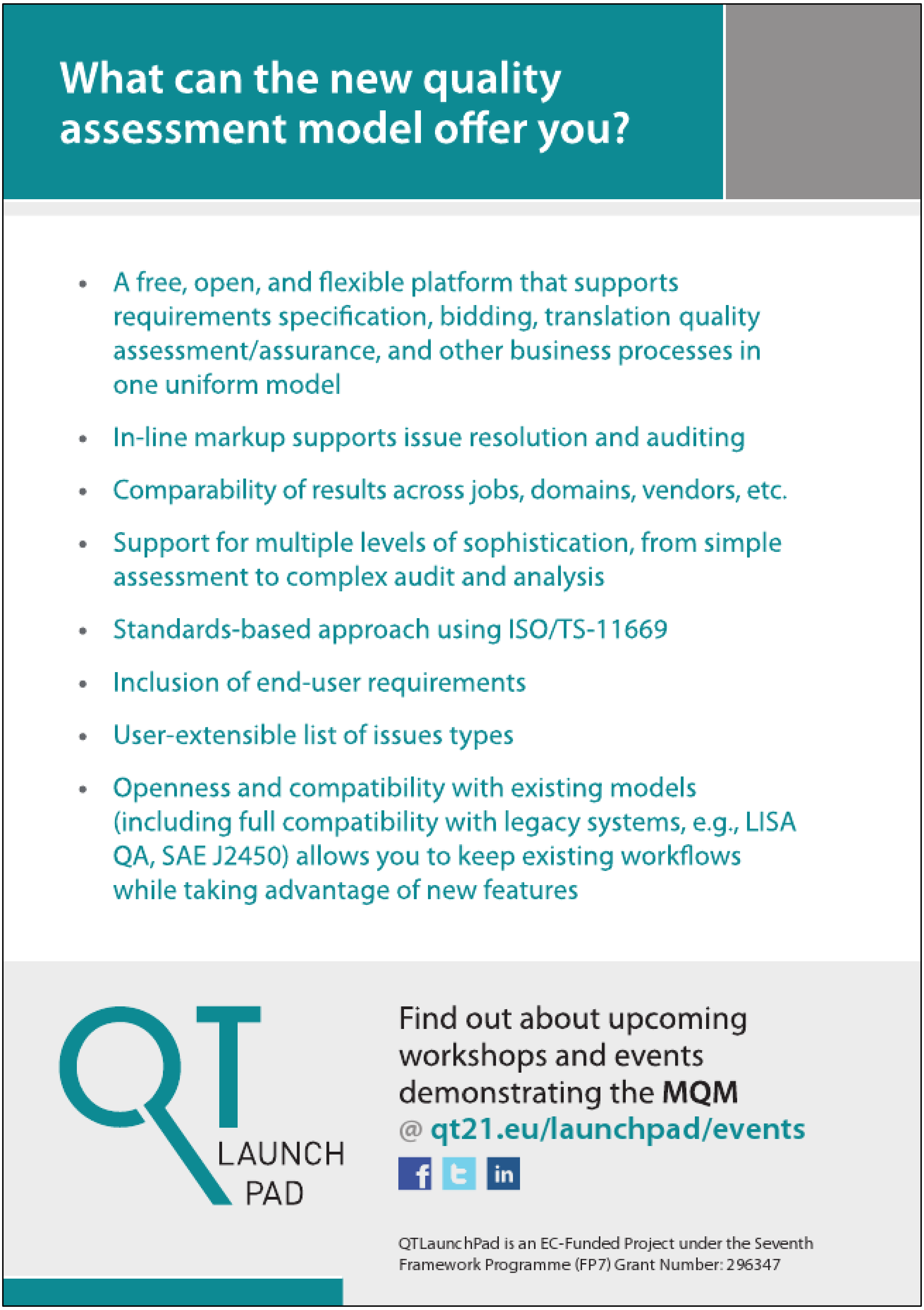
DFKI German Research Center for Artificial Intelligence

The University Of Sheffield

IEA ILSP

DCU

SEVENTH FRAMEWORK PROGRAMME



What can the new quality assessment model offer you?

- A free, open, and flexible platform that supports requirements specification, bidding, translation quality assessment/assurance, and other business processes in one uniform model
- In-line markup supports issue resolution and auditing
- Comparability of results across jobs, domains, vendors, etc.
- Support for multiple levels of sophistication, from simple assessment to complex audit and analysis
- Standards-based approach using ISO/TS-11669
- Inclusion of end-user requirements
- User-extensible list of issues types
- Openness and compatibility with existing models (including full compatibility with legacy systems, e.g., LISA QA, SAE J2450) allows you to keep existing workflows while taking advantage of new features

QT LAUNCH PAD

Find out about upcoming workshops and events demonstrating the **MQM** @ **qt21.eu/launchpad/events**

QTLaunchPad is an EC-Funded Project under the Seventh Framework Programme (FP7) Grant Number: 296347

Appendix C - Mailing List Material

Message Begins

***

QTLaunchPad Workshops at MultilingualWeb, Rome

Dear Colleagues,

I wish to invite you to a workshop hosted by the EC-funded QTLaunchPad project on multidimensional quality metrics (MQM), and on use-cases for a large-scale future MT research initiative (RIAS), co-located with MultilingualWeb W3C in Rome, Italy.

Workshop Dates: March 14th, 2013

Time: MQM 09:00 - 12:00 (lunch included for both workshops); RIAS 13:00–16:00

Venue: Headquarters of the United Nations Food and Agriculture Organisation, Rome, Italy

Cost: Free

1. Workshop on Multidimensional Quality Metrics (MQM)

Translation Quality Assessment (TQA) has recently emerged as an important business topic where formal metrics such as the LISA QA Model and SAE J2450 for human translation have helped, but automatic metrics for machine translation are currently suitable only for research projects, not for production environments. QTLaunchPad has developed a unified multidimensional framework for TQA that is built around quality metrics that move beyond the limitations of existing models and focus on richness and compatibility with usability as a core feature.

This workshop focuses on the measurement of translation quality. It introduces attendees to the metrics introduced above. It will demonstrate tools for creating project type-specific metrics and ensuring their validity for actual production tasks. Participants will further be invited to provide feedback and to discuss their own quality requirements and needs to help improve and further develop the model in a discussion-oriented exploration of key issues related to TQA.

- *Click here to attend and find out more about this workshop*

2. Workshop on Research Innovation Application Scenarios (RIAS)

A central aim of QTLaunchPad is the preparation for a large-scale research and innovation action (QT21) in the application of research into of several core areas which have been identified in close consultation with stakeholders in research and industry. These research innovation application scenarios or RIAS represent promising combinations of tasks, domains, users, industrial actors, demonstrators, innovation mechanisms, data, etc. Current suggestions under discussion include:

- **Automotive:** technical documentation with the end user in mind, and internal communication for multilingual environments;
- **Medical**: high-quality medical information for the general public, and emergency warnings (e.g. earthquakes) when multilingual data (e.g. SMS) need to be disseminated quickly and accurately;
- **Public**: public consultations and information;
- **Media**: subtitling/audio descriptors, e.g. for lectures and person-to-person communication;
- **Language Learning**: multi-modal communications a foreign language that is mastered only partially.

This workshop presents the progress of the exploration of these areas to participants and invites interactive discussions where attendees can add their own needs and requirements and provide welcomed feedback to the work carried out so far.

- *Click here to attend and find out more about this workshop*

Thanks and best wishes on behalf of the QTLaunchPad team,

Stephen

***

Message Ends