online Übungssystem

Unicode-Support und dynamische Charset-Erkennung

Umstellung des Online-Übungssystems auf Unicode-Unterstützung und Verarbeitung unterschiedlich kodierter Texte

Version:
1.2, 08/09.10.2014

FernUniversität in Hagen

Inhaltsverzeichnis

Abstract

Das Online-Übungssystem wurde mit der Version vom 08.10.2014 umfangreichen Änderungen unterzogen. Sofern sich dabei keine Fehler eingeschlichen haben, sollten diese Änderungen jedoch praktisch unbemerkt bleiben, insbesondere soll alles weiterhin so funktionieren wie bisher, nur mit u.a. folgenden zusätzlichen Möglichkeiten:

- Studentische Eingaben in Klartext-Eingabefelder können ab sofort beliebige Sonderzeichen enthalten.
- Dasselbe gilt insbesondere für In-Browser-Korrekturen.
- Von Kursbetreuern extern bearbeitete und ins Übungssystem hochgeladene HTML-Dateien (z.B. Aufgabenformulare, Musterlösungen etc.) dürfen nun mit nahezu beliebigem Charset (z.B. `UTF-8`) gespeichert werden. Die Kodierungen von zu kombinierenden Dateien (Seitenvorlagen und darin einzubettende Dokumente) dürfen unterschiedlich sein und werden automatisch angeglichen.

Einleitung

Das Online-Übungssystem stammt von einem System namens *WebAssign* ab, das in den Anfängen der Webapplikationen, in den 90er Jahren entwickelt wurde. In der Zwischenzeit hat sich im Web sehr vieles getan, und auch am Online-Übungssystem fanden bereits in den zurückliegenden Jahren umfangreiche Änderungen statt, so dass das originale WebAssign-System kaum noch wiederzuerkennen ist. So wurde z.B. insbesondere das Webinterface grundlegend modernisiert und flexibilisiert. Für einfachere Aufgabenerstellung wurde ein neuer Erstellungsassistent für handbewertete Aufgaben eingeführt. Als Ersatz für die betagte CORBA-Schnittstelle zur Anbindung externer Korrekturlogik wurde eine neue SOAP-Schnittstelle in Betrieb genommen (wobei die CORBA-Schnittstelle aber aus Abwärtskompatibilitätsgründen vorerst noch weiter betrieben wird).

Ein spezielles Gebiet, auf dem sich im Web viel getan hat, ist die Zeichenkodierung für die Übersetzung von Texten in computerlesbare Bytefolgen (und umgekehrt). Früher wurden in der Regel einfache Zeichensätze verwendet, die den Nachteil eines mit 256 Zeichen relativ kleinen Alphabets hatten. In der Folge entstanden immer mehr verschiedene Kodierungen, die sich insb. in der Auswahl der darstellbaren Zeichen unterschieden – mit der Folge, dass einmal digital gespeicherte Texte beim Wieder-Einlesen häufig falsch interpretiert werden, wenn nicht zum Dekodieren wieder derselbe Zeichensatz verwendet wird wie für die Kodierung. Wer kennt sie nicht, die Probleme, dass Umlaute oder andere Sonderzeichen in Texten wie E-Mails oder Webseiten „verkrüppelt“ angezeigt werden (d.h. irgendwelche scheinbar unsinnigen Symbole an ihrer Stelle stehen)?

Heutzutage gibt es mit Unicode <http://de.wikipedia.org/wiki/Unicode> zumindest einen umfangreichen und wachsenden Zeichenvorrat, die (De-)kodierungsprobleme gehören damit aber immer noch nicht ganz der Vergangenheit an – mit den verschiedenen Unicode-Textkodierungen existieren ja nun sogar noch mehr verschiedene Kodierungen (neben den „althergebrachten“).

WebAssign und in Folge das Übungssystem gingen bisher mit Zeichenkodierungen eher „naiv“ um und stellten gewisse Anforderungen an die Ausführungsumgebung (Server, Datenbanken) und Kodierung der hinterlegten Dokumente und Informationen. Nun wurde das Übungssystem in dieser Hinsicht umfangreich modifiziert. Die Modifikationen dienen teils interner Optimierung (mehr Geschwindigkeit, weniger Speicherbedarf), Flexibilisierung und besserer Zukunftssicherheit (z.B. Voraussetzung für Datenbankumstellungen), teils aber auch der Erweiterung des Funktionsumfangs wie in den folgenden Abschnitten genauer vorgestellt.

Unicode-Support

Bisher nutzte das Online-Übungssystem weitestgehend die früher üblichen 8-Bit-Zeichensätze `ISO-8859-1` (alias „Latin-1“) bzw. dessen „Erweiterung“ `Windows-1252` für Webseiten und `ISO-8859-15` (alias „Latin-9“) für Datenbankeinträge wie Aufgabennamen.

Für ausgewählte Daten (die in binär in der Datenbank gesichert werden und somit unabhängig vom Datenbanksystem beliebig kodiert werden können) wurde nun ein Unicode <http://de.wikipedia.org/wiki/Unicode> -Support eingeführt. Dieser hat insbesondere Auswirkungen auf die folgenden Bereiche:

Einsendungen / Korrekturen

Insbesondere Klartext-Einsendungen von Studenten erfolgen ab sofort Unicode-basiert: Wann immer Sie in einer Aufgabe ein Plaintext-Eingabefeld für studentische Eingaben (einfache Eingabezeilen oder mehrzeilige Textareas ohne WYSIWYG-Editor[1] für formatierten Text) eingebunden haben, können Studenten ab sofort beliebige Unicode-Sonderzeichen einsenden, während früher Eingaben wie z.B. ≤, ≠, ∞, ø oder auch einfache typographische Anführungszeichen („“) in der Regel nicht verlustfrei eingesendet werden konnten.

Dasselbe trifft auch auf Korrektor-Kommentare im Rahmen einer In-Browser-Korrektur zu.

Korrekturmodule kommen mit Unicode-Einsendungen klar, sofern sie über die neue SOAP-Schnittstelle angebunden wurden und nicht mehr über die alte CORBA-Schnittstelle arbeiten. Letztere konnte aus Abwärtskompatibilitätsgründen zu bestehenden Korrekturmodulen nicht mehr angepasst werden und überträgt Unicode nicht verlustfrei.

HTML-Ressourcen

Im Übungssystem hinterlegen Kursbetreuer / Aufgabenautoren in der Regel Inhalte als HTML-Dateien. Im einfachsten Fall lassen Sie die Aufgaben assistentengestützt generieren und bearbeiten die generierten Dateien auch nicht manuell nach. Wenn Sie aber von Hand HTML-Dateien erstellen oder generierte Dateien herunterladen, offline nachbearbeiten und wieder hochladen, dann mussten Sie sich bisher in der Regel auf das Encoding `Windows-1252` oder `ISO-8859-1` alias `ISO-Latin-1` beschränken. Bei anders kodierten Dateien hätte zumindest der Embedding-Mechanismus (also die Markierung des eigentlichen Inhalts über Marken `$EMBED` und `$/EMBED`, so dass der so markierte Inhalt in eine zentral hinterlegte Seitenlayout-Vorlage eingebettet wird) nicht mehr korrekt funktioniert.

Diese Beschränkung ist nun gefallen, nun dürfen Sie Ihre Dateien (Aufgabenstellungen, Musterlösungen etc.) insbesondere auch in Unicode-Encodings wie typischerweise `UTF-8` speichern und ins Übungssystem hochladen. Tatsächlich steht Ihnen die Zeichensatz-Wahl weitgehend frei, vgl. auch folgenden Abschnitt.

CSV-Exporte

Als Betreuer können Sie verschiedene Daten im CSV-Format exportieren, z.B. Studientagsanmeldungen oder Leistungstabellen. Diese CSV-Dateien werden ab sofort im UTF-8-Encoding erstellt.

E-Mailversand

Die Mailing-Funktionen, insb. Serien-E-Mails eines Betreuers an alle oder bestimmte Beleger, erlaubt nun ebenfalls beliebige Sonderzeichen im Nachrichtentext, indem der Mailinhalt und -Betreff UTF-8-kodiert versendet werden.

Einschränkungen

Der Unicode-Support erstreckt sich derzeit noch nicht auf *alle* Zeichenfolgen im Übungssystem – schon da die derzeit verwendete Datenbank ihre Textfelder nicht unicode-basiert speichert[2]. In bestimmten Feldern wie Aufgabennamen werden daher derzeit noch nicht beliebige Zeichen verarbeitet. Falls Sie z.B. nach einer Umbenennung einer Aufgabe in deren Namen Fragezeichen (`?`) an Stelle bestimmter Sonderzeichen sehen, so wird das an dieser Stelle zuvor eingegebene Sonderzeichen vorerst nicht unterstützt.

Falls Sie Korrekturmodule verwenden, gilt außerdem die Einschränkung, dass die Korrekturmodul-*Eigenschaften*[3] aus Abwärtskompatibilitätsgründen zu bestehenden Korrekturmodulen weiterhin `Windows-1252`-kodiert sind und folglich nicht beliebige Unicode-Zeichen enthalten können.

Charset-Erkennung und -Kombinierbarkeit

Das Online-Übungssystem versucht nun, die genaue Kodierung von Zeichenfolgen (und damit den verwendeten Zeichensatz) möglichst exakt zu erkennen. Wenn Sie als Betreuer / Aufgabenautor z.B. HTML-Dateien für Aufgabenformulare, Musterlösungen etc. extern erstellen und ins System hochladen, steht Ihnen ab sofort die Wahl der Zeichenkodierung weitestgehend[4] frei. Sofern Ihre HTML-Datei das verwendete Charset deklariert (z.B. über `<meta charset="utf-8">` oder `<meta http-equiv="Content-Type" content="text/html; charset=ISO-8859-1">`[5] im HTML-Kopf), wird das Online-Übungssystem diese Deklaration in der Regel finden und auswerten. Findet es keine, wird es versuchen, den Zeichensatz zu erraten (unter der Annahme, dass bestimmte Sonderzeichen wie Umlaute, Euro-Symbol, typographische Anführungszeichen etc. in deutschen Texten viel wahrscheinlicher vorkommen als die „kryptischen" Sonderzeichen, die bei falscher Interpretation meist an deren Stelle erzeugt werden).

Auch wenn Sie den Embedding-Mechanismus (s.o.) verwenden und z.B. eine UTF-8-kodierte Datei hochladen, deren mit `$EMBED` und `$/EMBED` markierter Inhalt in eine Seitenvorlage eingefügt werden soll, die ihrerseits z.B. „nur" Latin-1-kodiert ist, so kann das Übungssystem nun von beiden Dateien die jeweilige Kodierung erkennen, beide geeignet zusammenfügen und für das Ergebnis selbständig wieder eine passende Kodierung (hier: UTF-8) wählen, um alle Inhalte verlustfrei an den Browser ausgeben zu können.

Die Seitenkodierung Ihrer Aufgabenseiten hat nun auch keine Auswirkungen mehr darauf, wie die studentischen Formular-Eingaben interpretiert werden. Sie sind also völlig frei und können Aufgabenformularseiten bedenkenlos z.B. Latin-1-kodiert speichern, und dennoch werden die studentischen Einsendungen in Unicode (UTF-8) übertragen und verarbeitet[6].

Fußnoten

1. Eingabeboxen, die mit dem HTML-Editor „TinyMCE“ ausgestattet sind, ließen auch in früheren Übungssystem-Versionen bereits beliebige Sonderzeichen zu. Diese wurden vom TinyMCE automatisch zu HTML-Ersatzdarstellungen (sog. Entites) konvertiert.
2. Die mit dieser neuen Übungssystem-Version vorgenommenen Änderungen schaffen jedoch überhaupt erst die Voraussetzung dafür, dass die Datenbank in Zukunft ebenfalls auf ein Unicode-Format wie `UTF-8` umgestellt werden kann.
3. Korrekturmodul-Eigenschaften: Ein in der erweiterten Aufgabenerstellung zu einer Aufgabe hinterlegter Text aus Schlüsselnamen und Werten, der Optionen aufnimmt, die zur konkreten Aufgabe ans Korrekturmodul übergeben werden, z.B. Einstellung der erreichbaren Punkte (für solche Eingenschaften spielt diese Charset-Einschränkung keine Rolle), möglicherweise aber auch Kommentartexte, die in gewissen Fällen automatisch in die erzeugte Korrektur eingefügt werden sollen. In letzteren sind Nicht-Windows-1252-Sonderzeichen entweder zu vermeiden oder durch HTML-Entities zu kodieren (sofern die Korrekturmodul-Ausgabe im HTML-Format und nicht als Plaintext erfolgt).
4. Möglicherweise wird das Übungssystem nicht alle „Exoten“ unter den existierenden Zeichensätzen erkennen, aber die üblichen Charsets wie insbesondere `ISO-8859-1`, `ISO-8859-15`, `Windows-1252`, `UTF-8`, sogar `MacRoman`, `UTF-16` oder `UTF-32` werden unterstützt. Von den eben genannten abweichende Charsets werden höchstens bei Deklaration in Meta-Tags unterstützt, jedoch *nicht* von der automatischen Erkennung „geraten“.
5. Besonderheit: Genau wie Webbrowser wird auch das Online-Übungssystem Deklarationen von `ISO-8859-1` intern als `Windows-1252` auswerten. Beide Zeichensätze unterscheiden sich darin, dass `ISO-8859-1` spezielle Steuerzeichen enthält, die im Web nie verwendet werden, während `Windows-1252` die Positionen dieser Steuerzeichen statt dessen durch zusätzliche sichtbare Sonderzeichen wie z.B. die typographischen Anführungszeichen („“) oder das Euro-Symbol (€) belegt. In der Praxis finden sich tatsächlich häufig `Windows-1252`-kodierte Dateien, die von sich selbst „behaupten“, `ISO-8859-1`-kodiert zu sein, obwohl sie Sonderzeichen enthalten, die in `ISO-8859-1` gar nicht definiert sind, wohl aber in `Windows-1252`.
6. Das Übungssystem fügt zu diesem Zweck eigenständig ein `accept-charset`-Attribut ins Formular-Tag der Aufgabenseiten ein. (Voraussetzung ist, dass Sie ein solches nicht bereits selbst in Ihre Aufgabenformular-HTML-Datei eingetragen haben!)